Technical Report
824

# Floating Gate Circuits in MOSIS

J.R. Mann



1 November 1990

Lincoln Laboratory
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
LEXINGTON, MASSACHUSETTS

Prepared for the Defense Advanced Research Projects Agency
under Contract F19628-90-C-0002.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
LINCOLN LABORATORY

# FLOATING GATE CIRCUITS IN MOSIS

*J.R. MANN*
*Group 23*

TECHNICAL REPORT 824

1 NOVEMBER 1990



LEXINGTON MASSACHUSETTS

## ABSTRACT

The MOSIS foundry offers a two-poly CMOS process that can be used as a floating gate technology, albeit not with the same performance as commercial EEPROM foundries. This report characterizes the structures and programming techniques necessary to build floating gate structures and associated high-voltage addressing circuitry on the low-noise analog process available through MOSIS. Techniques that are used include Fowler-Nordheim tunneling, channel hot-electron injection, and avalanche injection. The dielectric materials between the floating gate and both the control gate and substrate are characterized. Unconventional lightly doped drain FET devices and additional circuit techniques for handling the high-voltage programming signals are presented.

# TABLE OF CONTENTS

# LIST OF ILLUSTRATIONS

## LIST OF ILLUSTRATIONS
## (Continued)

# 1. INTRODUCTION

Since MOSIS first announced the addition of a double poly process to its standard fabrication runs, it has been widely recognized that the possibility exists for its use as a floating gate technology. Floating gates are useful not only as electrically erasable/programmable read-only memories (EEPROM), but also for electrically programmable logic devices (EPLD), as analog trimming devices [1], and as modifiable synaptic connections in neural networks [2,3]. Until now, the only technique available to MOSIS users for storing and removing charge from a floating gate has been ultraviolet radiation [4]. This is extremely slow, cannot be performed in situ, and requires shielding over active circuitry, thereby limiting its overall usefulness.

This report describes the circuit elements required to build an EEPROM using the MOSIS standard low-noise analog CMOS process. This process was selected because it includes additional implants that are critical to high-voltage circuitry. This technology is an enhanced N-well CMOS process that includes a P-base implant to realize NPN bipolar transistors and an N-type channel implant to make buried channel CCDs. Channel oxide thicknesses run around 400 $\mathring{A}$ while the interpoly oxide runs between 600 and 700 $\mathring{A}$ thick. This translates into capacitance values of $\approx 0.8 fF/\mu m^2$ and $0.5\ fF/\mu m^2$, respectively.

The first part of this report concentrates on the floating gate storage elements and the techniques used to program them. The latter part presents the devices that are used to generate and direct on-chip high-voltage signals. Following this is a review of the status of the circuit elements to build an EEPROM. It is our hope that a cell library can ultimately be made available to MOSIS users with tested circuit elements to realize a full EEPROM memory.

# 2. FLOATING GATE STORAGE ELEMENTS

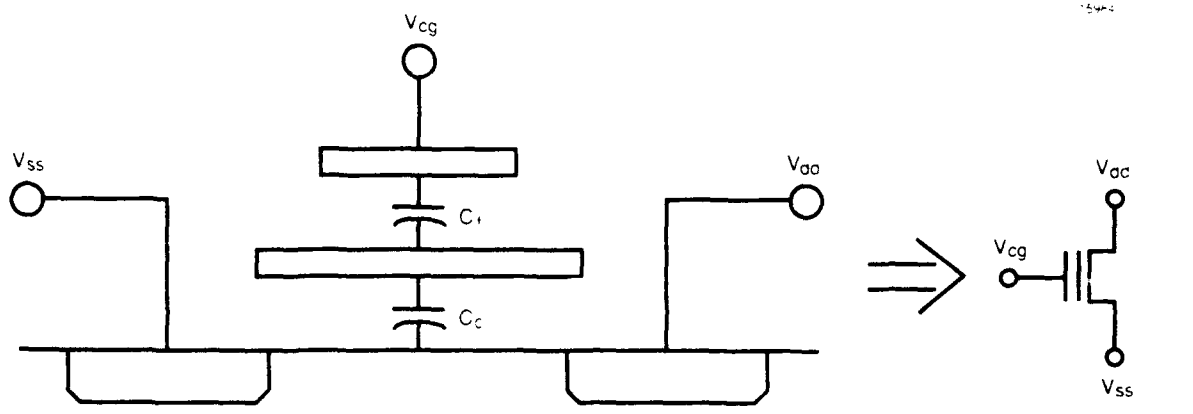
## 2.1 Method of Testing

Many techniques exist for depositing (writing) and removing (erasing) electrons from a floating gate. Each technique requires appropriate structures to establish the proper conditions to move this charge electrically. All the devices considered here have a double poly stacked gate arrangement where the first poly layer is sandwiched between the substrate surface and the second poly layer, called the control gate. This second layer induces the appropriate fields conducive to electron transfer to or from the floating (first level) poly gate. (Ultraviolet radiation to erase devices will not be considered here, though that may present a very worthwhile option to some users.) Three types of structures were considered: a simple stacked gate, an overlapped gate, and an avalanche injection, as illustrated in Figure 1.

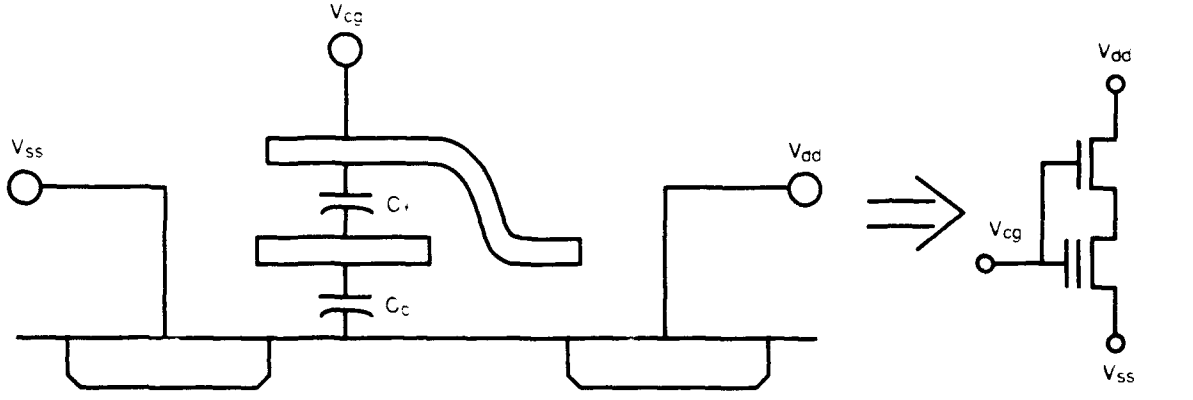
If carriers, accelerated in the high-field region near the drain, exceed the ambient thermal kinetic energy of the lattice they become "hot," allowing them to overcome the oxide energy barrier of 3.2 eV. In the presence of the appropriate fields at the drain end of the channel, these electrons are swept up and retained on the floating gate. This source of carriers is the channel current in hot-electron injection devices and is hole-electron pairs generated by impact ionization in the depletion region in avalanche injection devices. In contrast, Fowler-Nordheim tunneling is due to the increased quantum-mechanical tunneling probability of electrons through the oxide energy barrier with increased electric field strength.

Because the low-noise analog process does not include ultrathin oxides through which Fowler-Nordheim tunneling from the substrate can be accomplished, one of the hot-electron techniques must be used in writing these devices. However, asperities on the top surface of the floating poly layer do permit tunneling between the floating poly gate and the control gate by creating points of high electric field. This provides a mechanism for electrically erasing these devices. The stacked gate structures are written by hot-electron injection and erased by Fowler-Nordheim tunneling. The overlapped structures are used in flash EEPROM [5] where the series control gate acts as an access transistor and prevents conduction due to overerasure of the floating gate. It can be written either by channel hot-electron or avalanche injection (depending on gate length) and erased by Fowler-Nordheim tunneling. The avalanche structure is explicitly designed to be written by avalanche injection through a sourceless transistor (gated diode) and is also erased by Fowler-Nordheim tunneling.

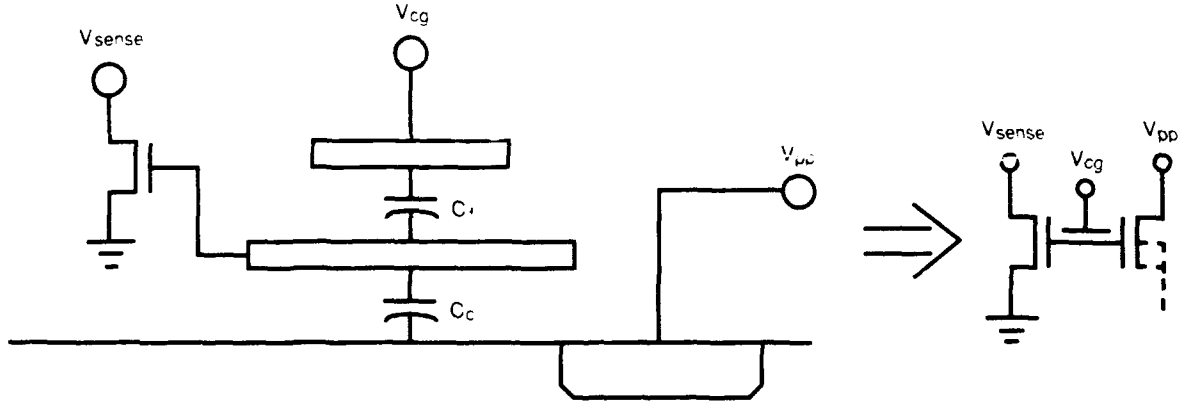
One critical design parameter for these devices is the ratio between the floating gate-to-control gate capacitance and the floating gate-to-channel capacitance. This determines how well the floating gate is coupled to the control gate, hence the division of the applied control gate potential ($V_{cg}$). On one hand, good coupling is needed to produce strong electric fields between the floating gate and the channel, hence more efficient electron collection. On the other hand, it is necessary to be able to create high fields between the control and the floating gates to provide tunneling during erasure. Therefore, the test structures included several capacitance ratios expressed as area ratios




(a) STACKED GATE




(b) OVERLAPPED GATE




(c) AVALANCHE

*Figure 1. Floating gate structures: (a) stacked, (b) overlapped, (c) avalanche.*

(e.g., 1:4 means the drawn control gate overlap of the floating gate is 1.4 the drawn area where the floating gate overlaps the thin oxide in the channel region). Various channel lengths were also included on hot-electron written devices in order to determine their limits.

The following technique has been employed to characterize the writing and erasing of the device under test. In both instances, one of the device terminals was pulsed with a certain voltage for a fixed time period, while the others were held at prespecified levels. The threshold voltage of the device was then measured, recorded, and compared with the previous threshold. If the *change in threshold* from the previous measurement fell outside a user-specified boundary, the cycle was repeated. If not, the device was restored to its starting value and a new cycle begun with a different set of parameters. For example, during a test for writing behavior, the device would be erased to an initial reference threshold voltage, then written under the conditions specified for this iteration. The threshold change from one pulse sequence to the next would diminish with the charging of the floating gate. If the rate of threshold change fell below a boundary specified by the user at the beginning of the test, the writing sequence would be terminated and the device erased to its initial threshold voltage in preparation for the next test. Erase testing followed essentially the same procedure, only the device was first written to an initial threshold level, then erased until the threshold change criterion was met. The device threshold voltage was always limited to some prespecified range (typically $\pm 8$ V) throughout the testing to prevent possible device damage.

The threshold voltage of a device was measured by using a binary search for a given drain current. The search began at 0-V $V_{gs}$ and worked outward, first in 1-V steps, then dividing the stepsize by 10 until a 1-mV precision was reached. The threshold current level was determined by multiplying the width-to-length ratio of the device under test (DUT) by $-0.5$ $\mu$A for P-type devices and 1 $\mu$A for N-type devices. This search was bounded by upper and lower limits set by the operator, typically $\pm 8$ V. If, during writing or erasing, the threshold measure of the device fell outside this range, no additional programming voltages were applied that might move the threshold any further in that direction. This automatically ensured that the threshold gradient would go to 0, stopping any test that might be in progress. An artifact of this approach is that the time recorded for reaching the end of a test will mean two different things depending on whether the test was terminated by this limit or by reaching a true maximum or minimum threshold. In the former case the times will indicate how long it takes to shift the threshold by a fixed amount, whereas the latter will indicate the time to reach a maximum/minimum value, the magnitude of which will be variable depending on the programming conditions. A good rule for interpreting the plots is to look at the recorded threshold voltages where the maximum or minimum limits were not reached and to observe the elapsed times on tests that reached these limits.

Erasing by Fowler-Nordheim tunneling was accomplished by pulsing the control gate ($V_{cg}$) to high positive voltages while both the drain and source were grounded. Hot-electron injection writing was accomplished by pulsing the drain while holding the control gate at a prespecified level, the source being held at ground. Avalanche injection writing occurred by applying a programming voltage ($V_{pp}$), the magnitude of which was greater than the junction breakdown voltage of the gated diode, while holding the control gate at a prespecified value. In series with the programming

supply voltage was a large resister (100 MΩ) to limit the current. Here the source and drain of the sensing transistor were grounded as they were not involved in the writing process.

## 2.2 Results

Because of the number of degrees of freedom associated with characterizing these devices, the results published here have been selected to represent what was deemed the most relevant to potential users. These data may not satisfy all readers, but should provide enough information to give them a satisfactory starting point for their own investigations.

Just a brief warning and reminder that while the writing capability may improve with increasing capacitive coupling ratios, the erasing capability is deteriorating. One must consider both aspects of a device before selecting a geometry for any particular application.

### 2.2.1 Writing N-Type Stacked Gate Structures

N-type floating gate structures were, overall, the most resistant to being written by any technique described above. Only the very shortest channel lengths (2 $\mu$m) could be used for hot-electron injection, with some very slight success at 3 $\mu$m.

Figure 2 shows data collected for a standard writing test. The DUT had an 8- × 2-$\mu$m gate and a 1:4 capacitor area ratio (i.e., the control gate overlapped the floating gate with an area of 2 × 2 $\mu$m). Figure 2(a) illustrates the maximum threshold voltage achieved by writing this device with different applied drain voltages ($V_{ds}$) shown as a function of control gate voltage. Figure 2(b) shows the corresponding times required to achieve these threshold shifts. This device saw its maximum threshold shift for the condition of 7 V on the drain and 10 V on the gate. Correspondingly, it took about 20 ms to reach the maximum threshold voltage of −4.5 V, for a total 3.5-V threshold shift.

One should also note the decreasing maximum threshold shift after the 10-V $V_{gs}$ point. Here, erasing begins and the threshold voltage balances at a point where the number of written electrons balances with the number removed. After 11-V $V_{gs}$, erasing dominates the process and no threshold shift results.

The solid line in Figure 2(b) at 0.0001 s represents the unit pulse duration used during the test and is the minimum time represented on this plot. A point lying on this line can be interpreted as a failed write test where no threshold shift was recorded.

Figure 3 shows the same plots for a device with a 4- × 2-$\mu$m gate and a 1:2 capacitance ratio. The increased coupling of the control gate to floating gate capacitor greatly improves hot-electron writing over the previous device. All the drain voltages exhibit similar behavior. The important information here is the time it takes to program this device. The best programming times are obtained with 15 V on both the drain and gate, achieving a final threshold voltage of around 6 V in about 20 ms.

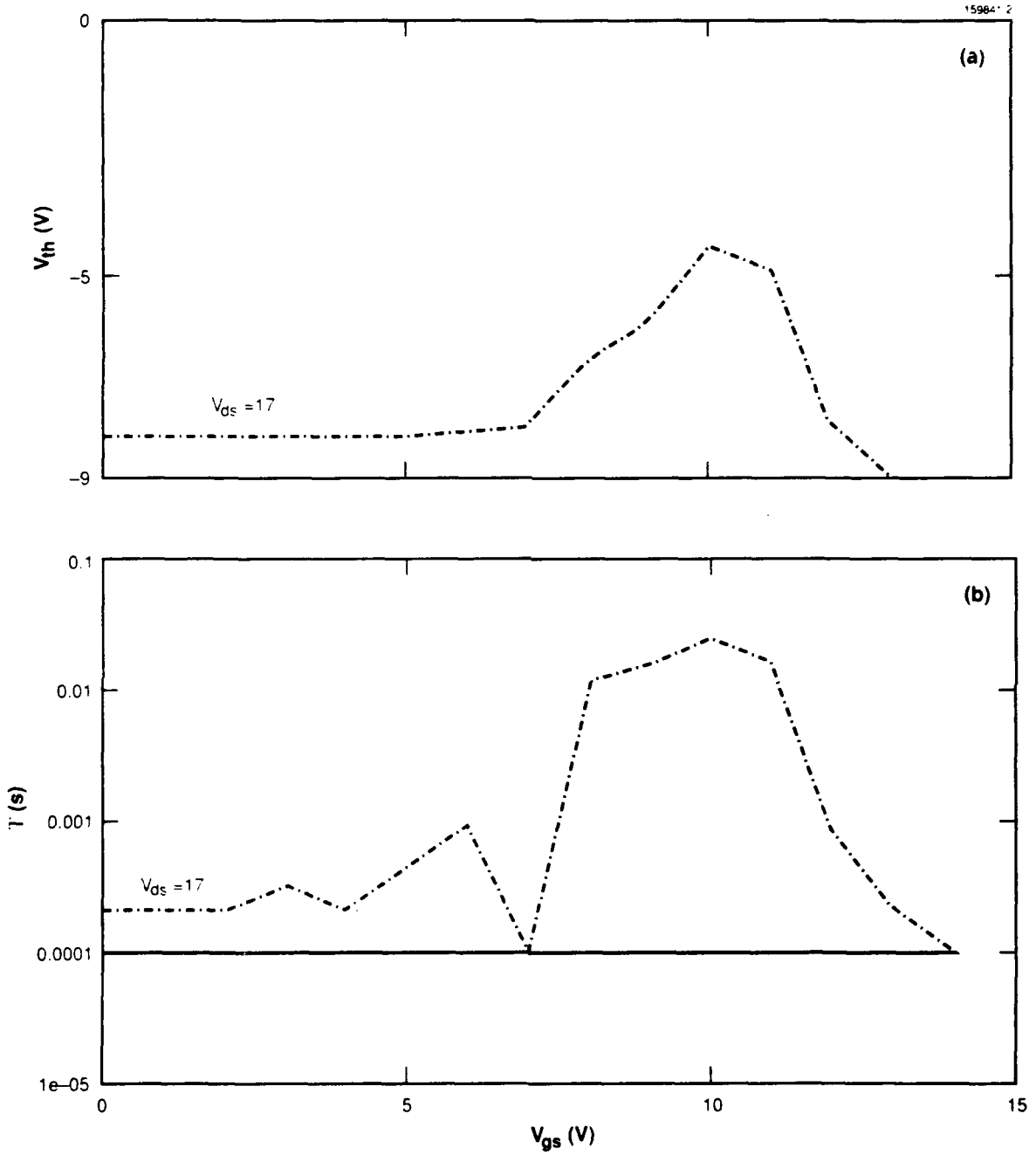



*Figure 2. N-type stacked gate structure, 1:4 coupling, 2-μm gate length: (a) maximum threshold voltage and (b) writing time.*

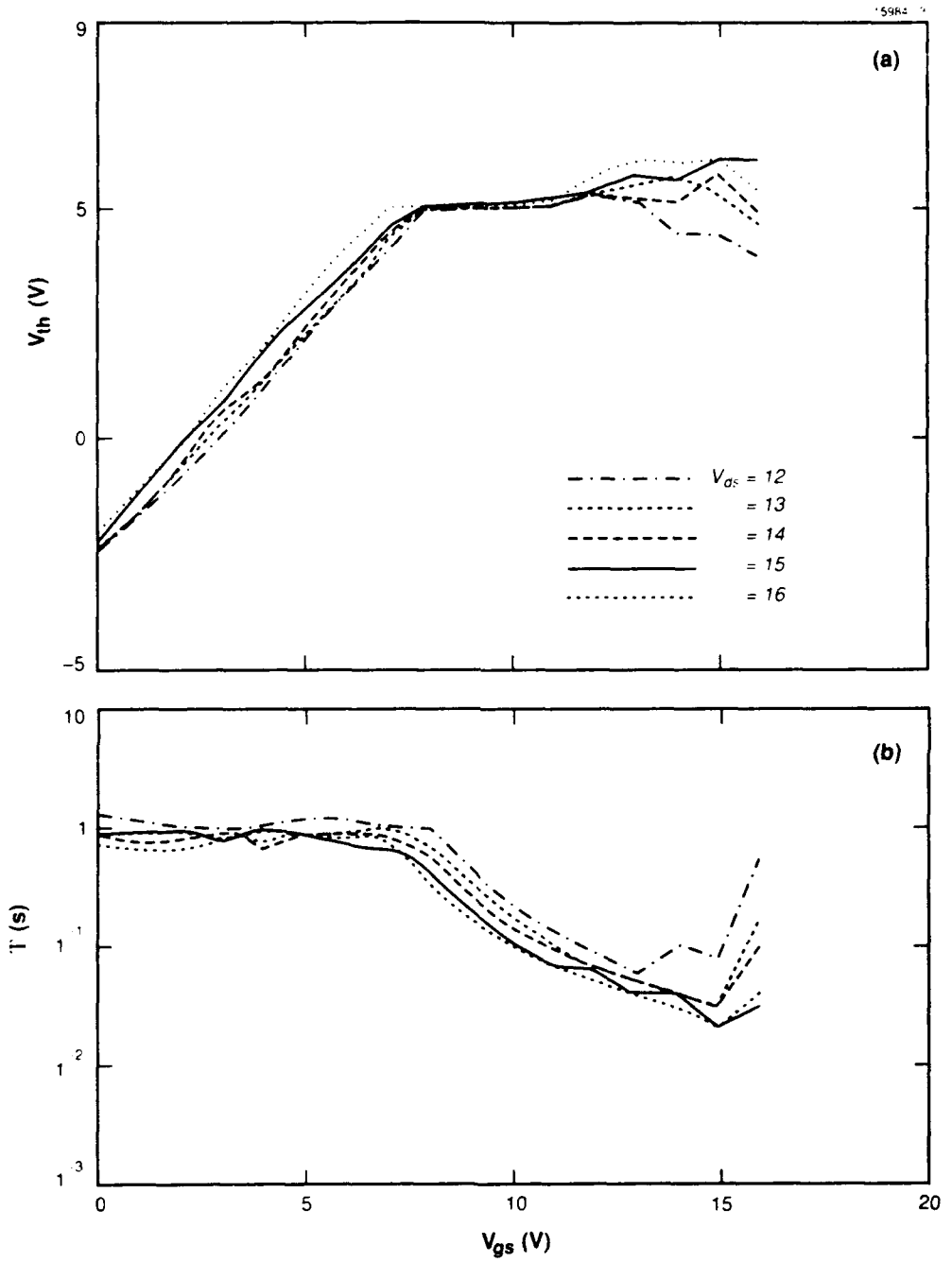



*Figure 3. N-type stacked gate structure, 1:2 coupling, 2-μm gate length: (a) maximum threshold voltage and (b) writing time.*

Figure 4 shows data taken for a device the gate of which was 4 × 2 $\mu$m with a 1:1 capacitive coupling ratio. Overall threshold shifts do not improve appreciably from the devices with lower coupling ratios, but the writing time is decreased significantly, down to 500 $\mu$s at $V_{gs}$ = 16 V and drain voltages at or above 12 V.

### 2.2.2 Writing P-Type Stacked Gate Structures

The P-type stacked gate structures have a unique programming characteristic: writing the device enhances the channel and produces higher channel currents; as a result, more carriers contribute to the gate current. This enhances the writing operation and should decrease programming time; however, erasing essentially makes it harder to turn the device on, to the point that it may exceed the writing gate voltage. Under this condition, no channel current would be generated and the device would not be written. In practice then, one would have to ensure that P-type stacked gate structures do not get overerased. Fortunately, there are writing conditions that can virtually guarantee that bits will not get stuck in this fashion. One should also note that the fields necessary to create the channel currents ($V_{gs}-$) are in opposition to those that draw electrons to the floating gate ($V_{gs}-$).

The plots in Figure 5 illustrate some of the above points. This device has an 8- × 2-$\mu$m channel and a 1:4 capacitive coupling ratio. Notice the sharp cutoff of the $V_{ds}$ = −9-V curve. After the gate voltage exceeds 3 V, the device will not conduct enough current to modify the threshold of this device, and in actual use would represent a stuck bit. Also notice that the same curve shows increasing threshold shift with increasing $V_{gs}$. This is due to the enhancement of the field attracting electrons to the floating gate. For both the −10 and −11-V $V_{ds}$, no bit sticking occurs. What distinguishes these conditions are the times necessary to reach these maximum thresholds. The −11-V $V_{ds}$ curve is about 10 times faster than the −10-V $V_{ds}$ condition. Also, increasing gate voltage also improves writing times over a factor of about 10. This device shows optimal performance at −11-V $V_{ds}$ and a gate voltage of 0 to 4 V.

Figure 6 shows similar data for a device with a 3-$\mu$m channel length. The −10-V $V_{ds}$ curve now more closely coincides with the −9-V $V_{ds}$ curve from the 2-$\mu$m device. Also, the times increase by a factor of 5 to 10 over the shorter device. A 4-$\mu$m length channel looks similar to the 3-$\mu$m curves except that the writing times increase by another factor of 2 to 4.

Figure 7 shows the data collected for a 2-$\mu$m channel length device with a coupling ratio of 1:2. Figure 8 shows plots for the same device with a 1:1 coupling ratio. One readily observable trend in the increased coupling is the tendency for the devices to get stuck at lower and lower gate voltages. At a 1:1 coupling, the maximum gate voltage is −5 V. Because of the increased coupling of the floating gate to the control gate, maximum attainable threshold voltages decrease for a given gate voltage. Also, for a given gate voltage the maximum attainable threshold shift for different drain voltages becomes smaller. At −10-V $V_{gs}$ on the 1:2 coupled device, the threshold voltages for the −9, −10, and −11-V $V_{ds}$ conditions are spaced about 6 V apart while the 1:1 coupled device shows a spacing of around 3 V.

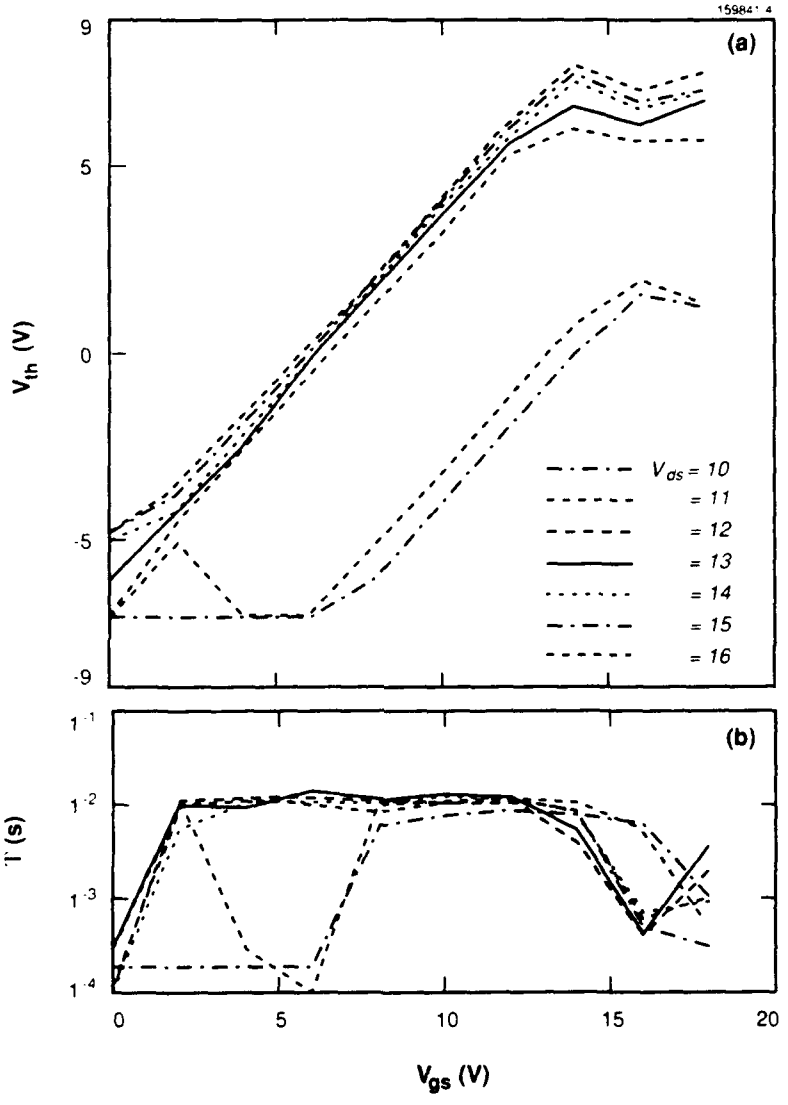



*Figure 4. N-type stacked gate structure, 1:1 coupling, 2-μm gate length: (a) maximum threshold voltage and (b) writing time.*

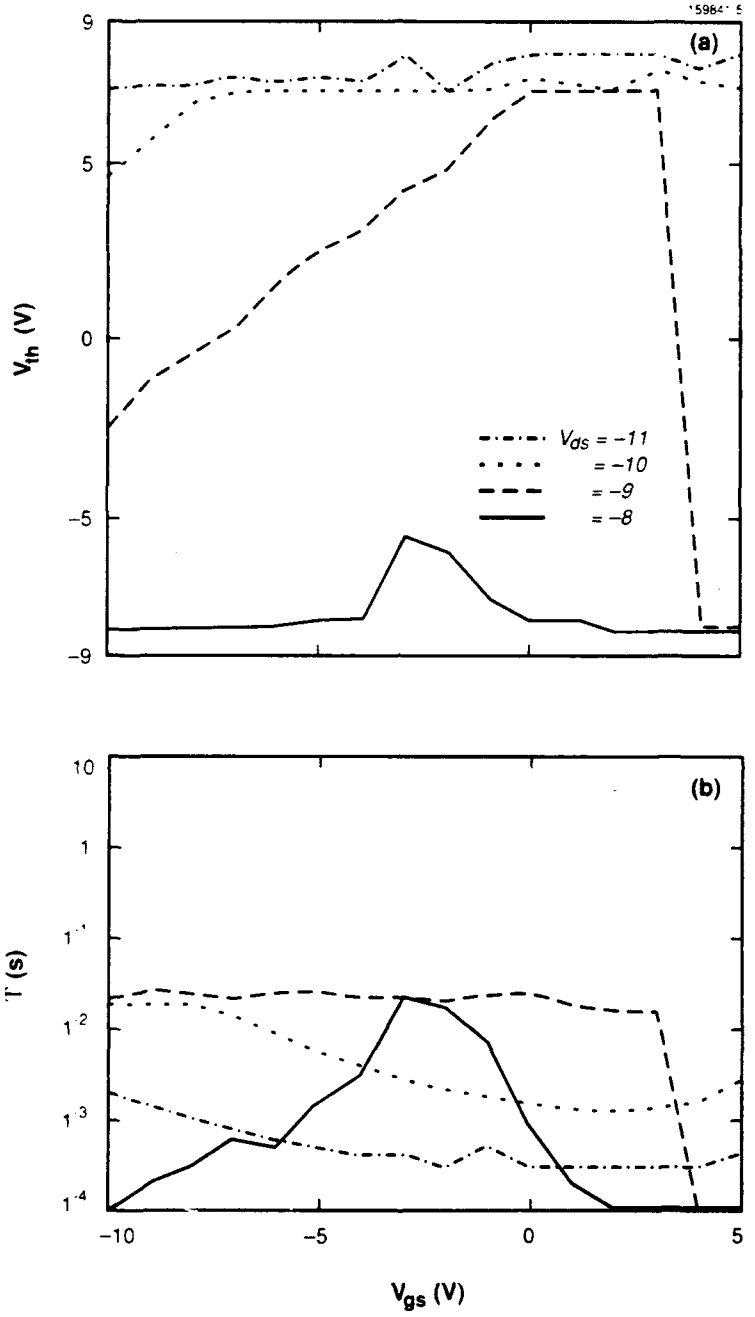



*Figure 5. P-type stacked gate structure, 1:4 coupling, 2-μm gate length: (a) maximum threshold voltage and (b) writing time.*

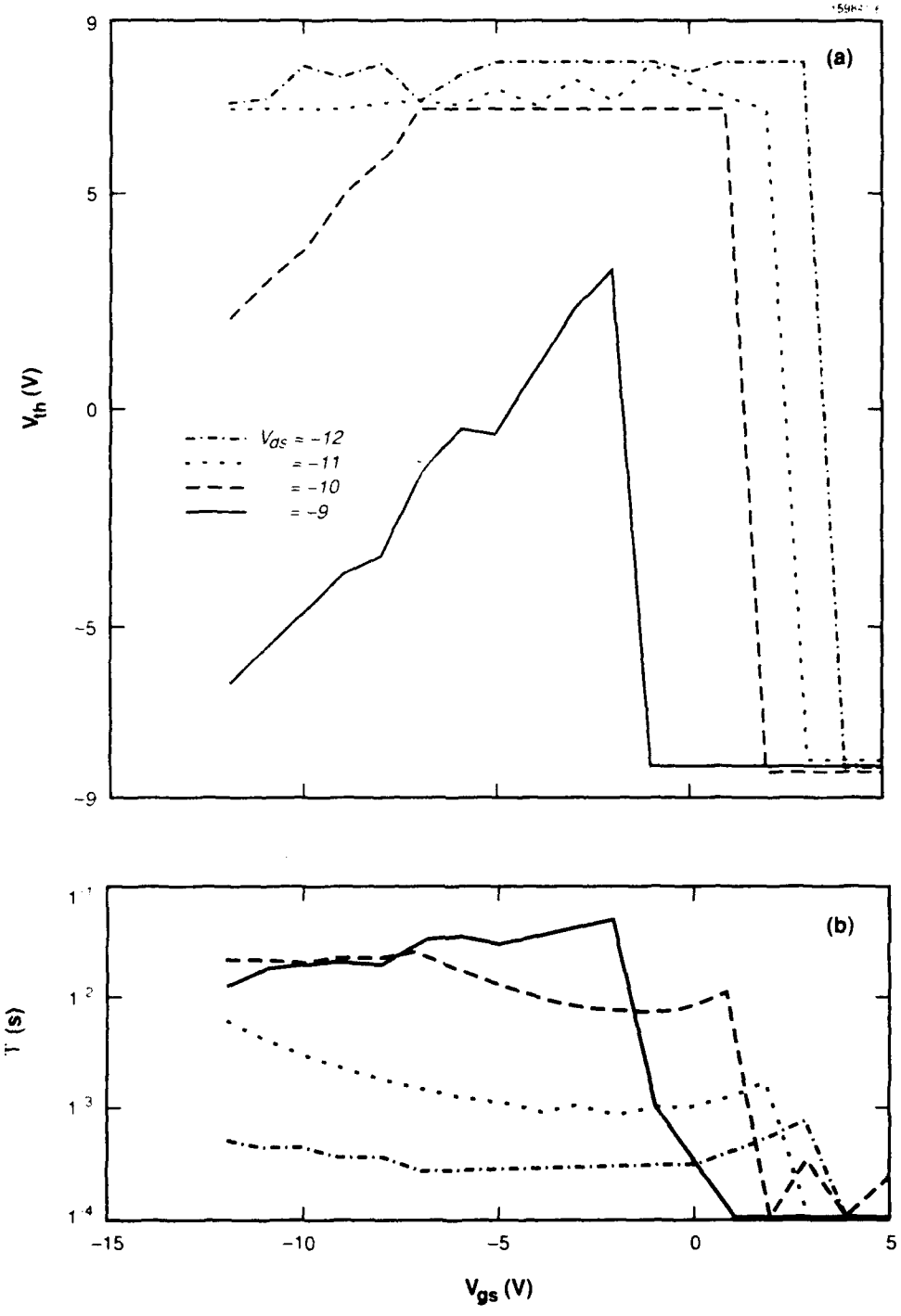



*Figure 6. P-type stacked gate structure, 1:4 coupling, 3-μm gate length: (a) maximum threshold voltage and (b) writing time.*

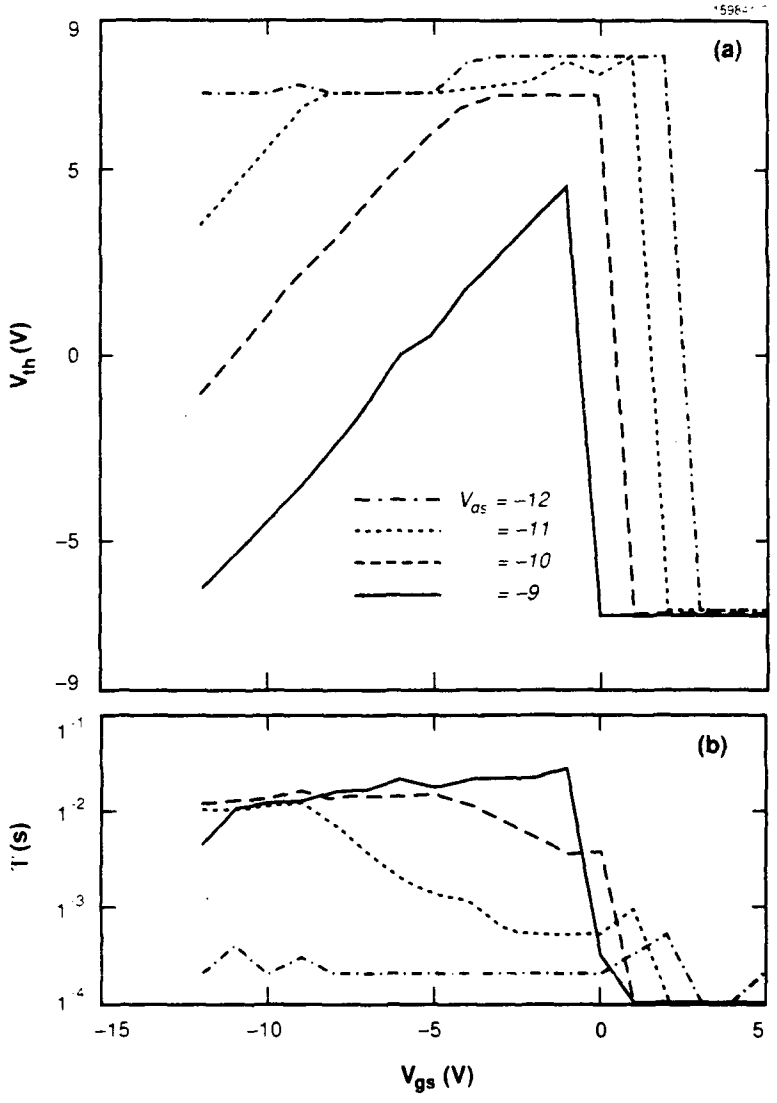



*Figure 7. P-type stacked gate structure. 1:2 coupling, 2-μm gate length: (a) maximum threshold voltage and (b) writing time.*

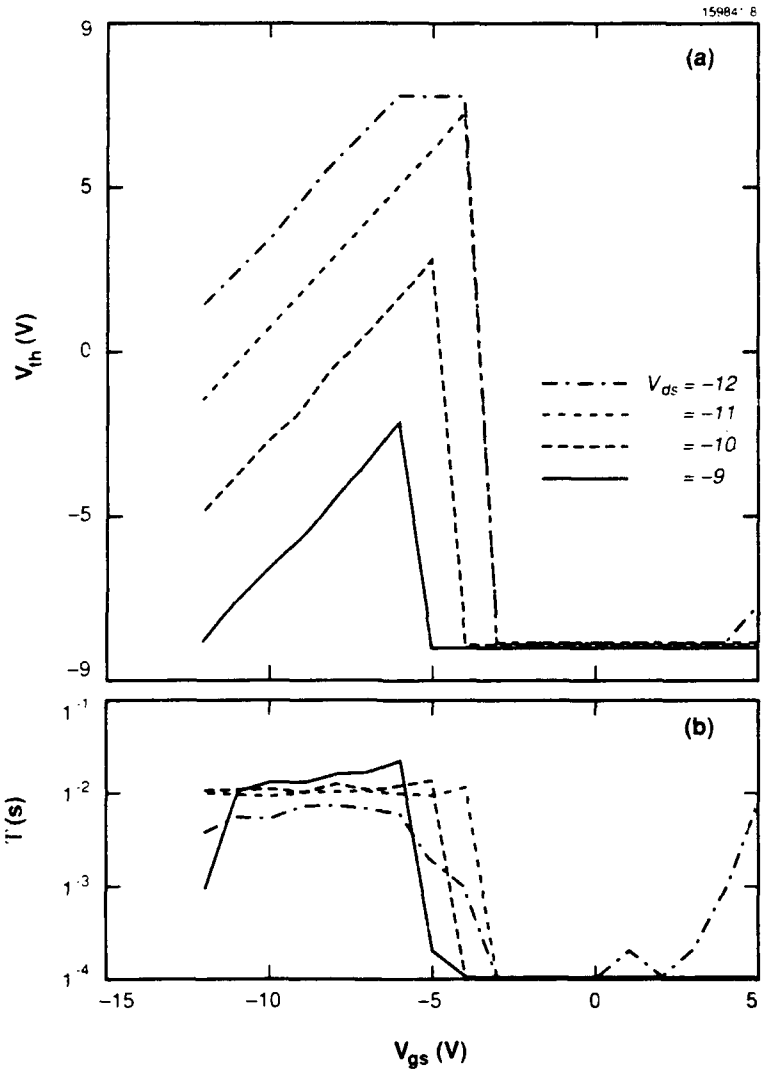



*Figure 8. P-type stacked gate structure, 1:1 coupling, 2-μm gate length: (a) maximum threshold voltage and (b) writing time.*

### 2.2.3 Writing N-Type Overlapped Gate Structures

Only one N-type overlapped gate structure could be written by channel hot-electron injection. This device had a floating and a control gate that were both 2 $\mu$m long, overlapped by 1 $\mu$m, leaving a total gate length of 3 $\mu$m. This obviously violates some design rules, but the device appeared to work satisfactorily. The results are shown in Figure 9. The coupling ratio on this device is 1:1. The designer is more restricted in deciding this coupling ratio for overlapped devices because a certain amount of overlap of the control and the floating gates is implicit in the structure.

The device could be written with a 12-V drain voltage and gate voltages of between 12 and 16 V. Increasing gate voltage reduced writing time by an order of magnitude every 3 to 5 V up to the point that erasing began to occur. The maximum threshold shift was about 4 V. Notice the effect of the series control-gated channel. The combined threshold voltage can never be less than the minimum of the two device thresholds in series. This guarantees that the floating gate device can always be turned off in the event that the floating gate itself is overerased. This is essential in memories built from floating gate devices permitting selective reading of memory bits connected to a common wire.

### 2.2.4 Writing P-Type Overlapped Gate Structures

The data shown in Figure 10 are for the same structure described immediately above, only P-type. Again, the series control-gated channel introduces a limit in the device threshold voltage. The curves look similar to those taken for a similar stacked gate structure except for the curves turning up below $-10$-V $V_{gs}$. At this point no explanation exists for this effect, but it is repeatable and appears in every overlapped P-type structure tested. Programming time for this structure was reduced to under a millisecond for a drain voltage of $-12$ V. This could probably be reduced further with more negative drain voltages.

The data in Figure 11 are taken from a device just like the one immediately above, only the control gate is 3 $\mu$m long and produces a more reasonable 2-$\mu$m control-gated channel and a total gate length of 4 $\mu$m. The curves are only slightly different than before with slightly lower maximum threshold shifts and longer writing times.

Data from an even more conservative design are given in Figure 12. This device is made from two 3-$\mu$m-long floating and control gates overlapped by 1 $\mu$m, leaving a total gate length of 5 $\mu$m. Now only drain voltages below $-10$ V are useful in writing these devices and writing times are a bit longer. Also, the maximum usable control gate voltage goes down.

The most conservative approach, where no design rules are violated, is composed of 4-$\mu$m-long floating and control gates overlapped by 2 $\mu$m, leaving an overall gate length of 6 $\mu$m. These data are plotted in Figure 13. The minimum drain voltage necessary to achieve maximum threshold shift is now $-11$ V. While these devices are even slower yet, more negative drain voltages can still reduce writing times to useful levels.

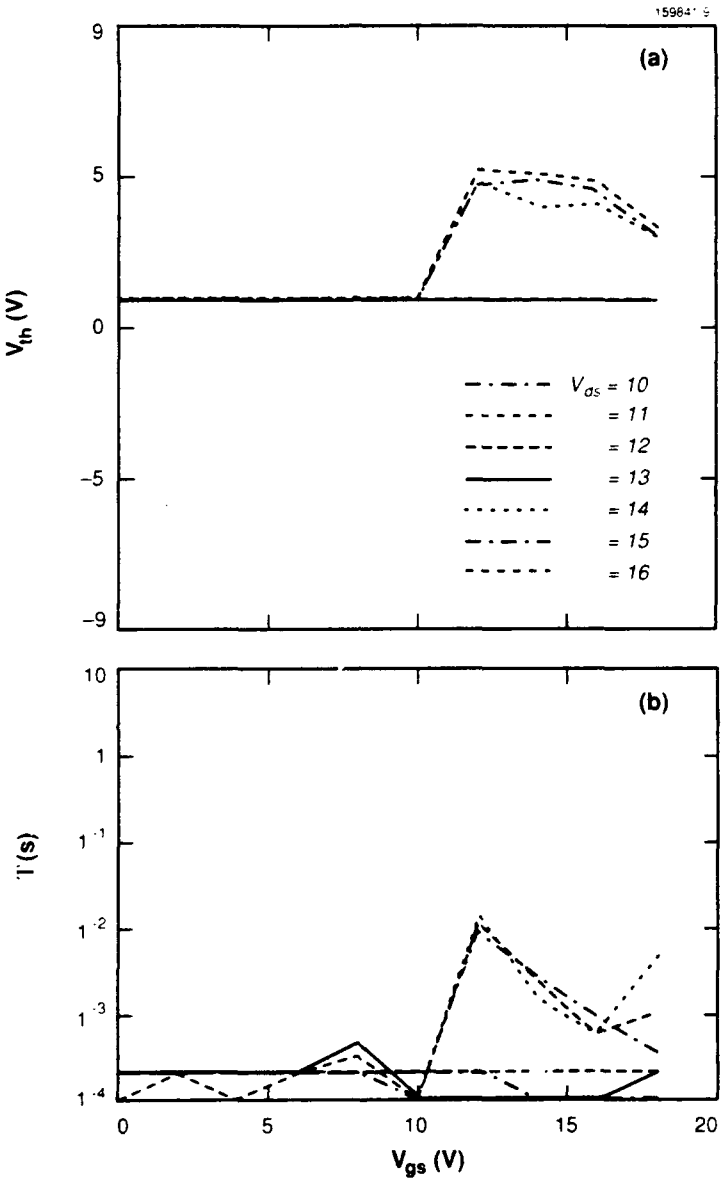



*Figure 9. N-type overlapped gate structure, 1:1 coupling, 2-μm floating gate, 2-μm control gate, 1-μm overlap: (a) maximum threshold voltage and (b) writing time.*

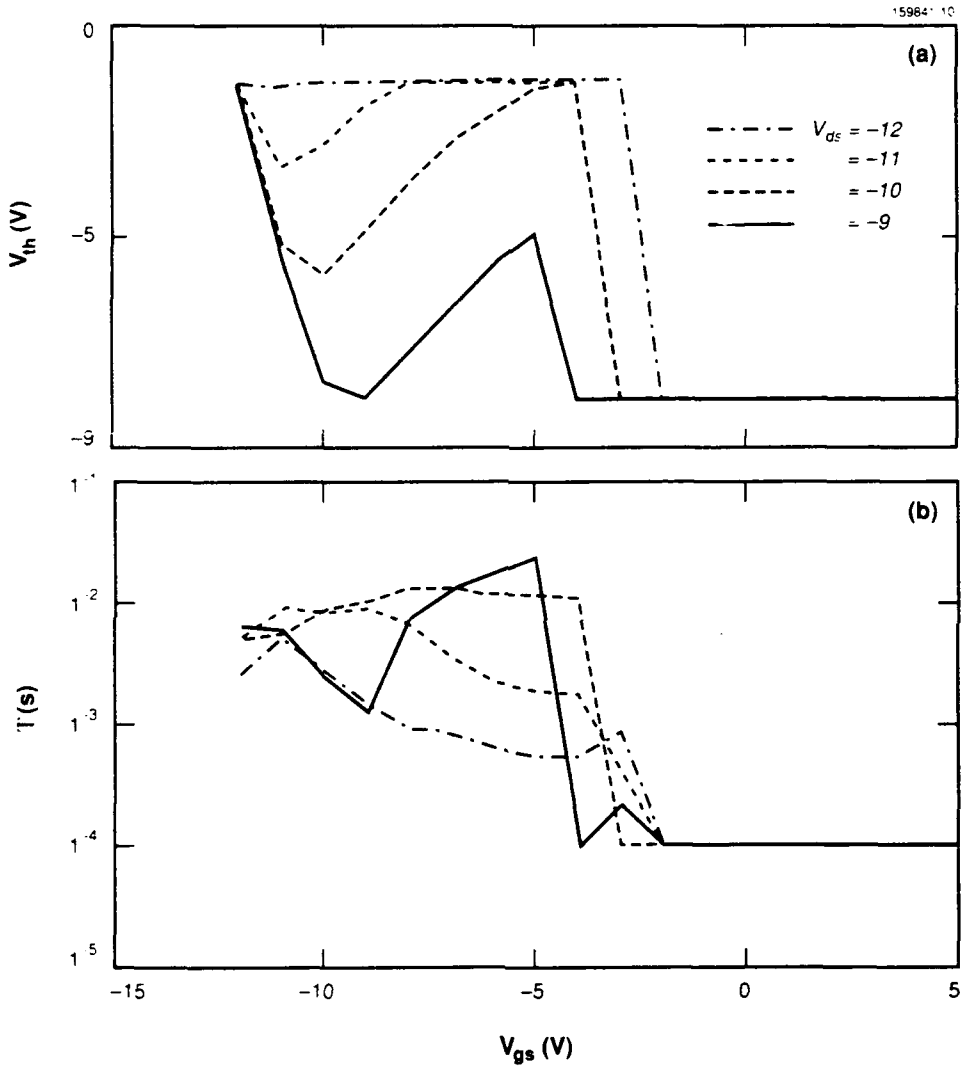



*Figure 10. P-type overlapped gate structure. 1:1 coupling. 2-μm floating gate. 2-μm control gate. 1-μm overlap: (a) maximum threshold voltage and (b) writing time.*

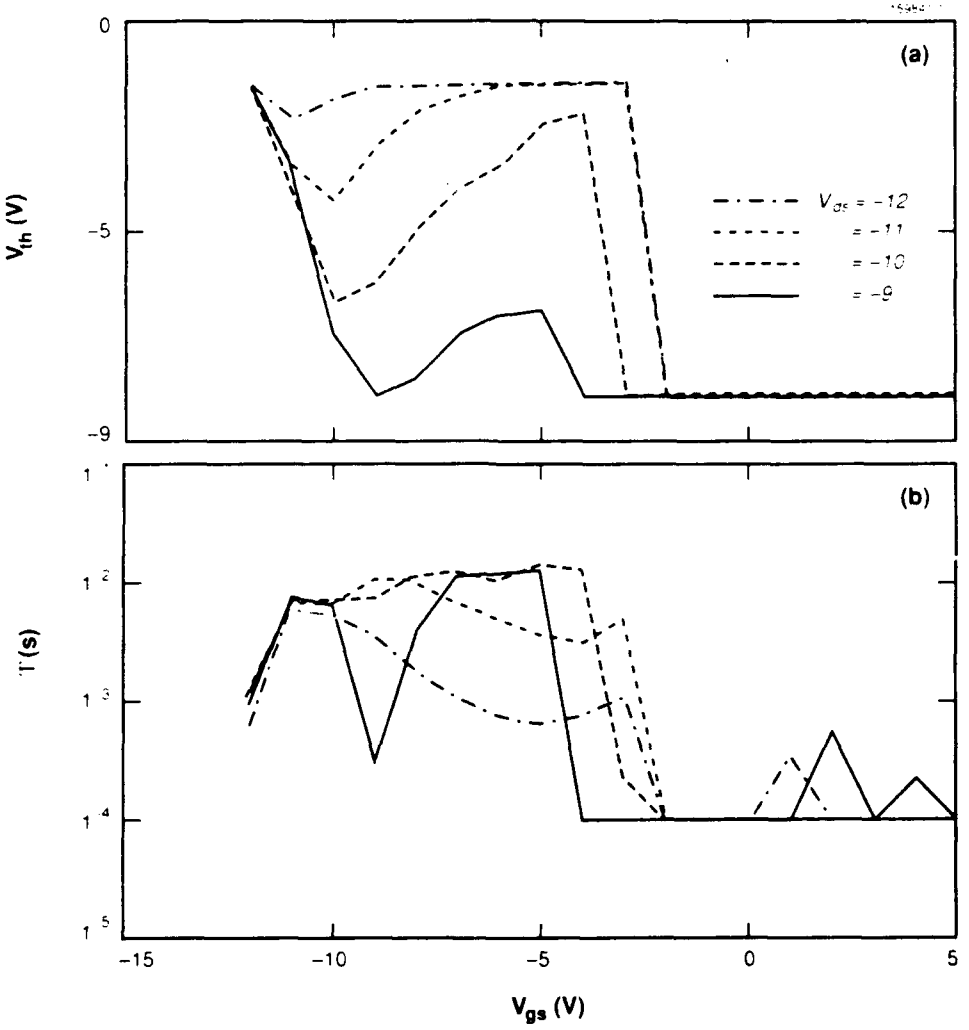



*Figure 11. P-type overlapped gate structure. 1:1 coupling. 2-μm floating gate, 3-μm control gate. 1-μm overlap: (a) maximum threshold voltage and (b) writing time.*

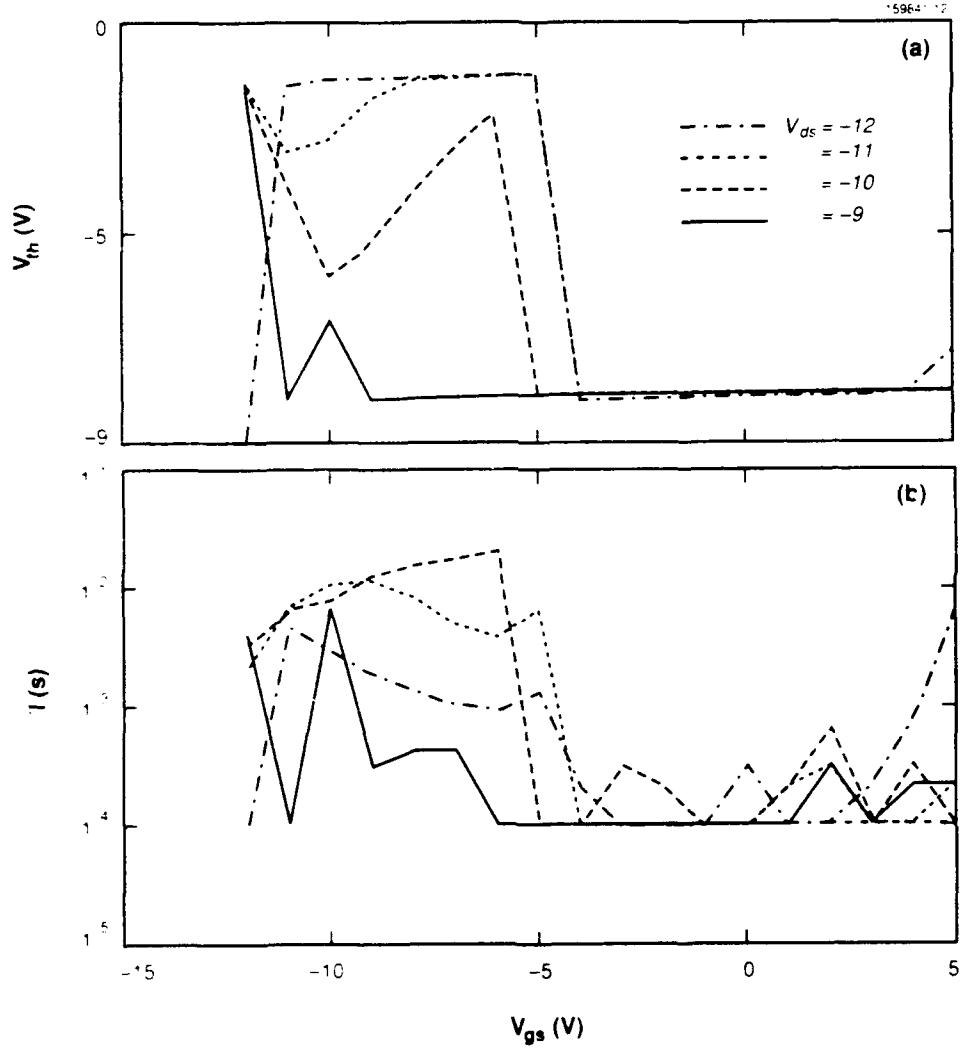



*Figure 12. P-type overlapped gate structure, 1:1 coupling, 3-μm floating gate, 3-μm control gate, 2-μm overlap: (a) maximum threshold voltage and (b) writing time.*

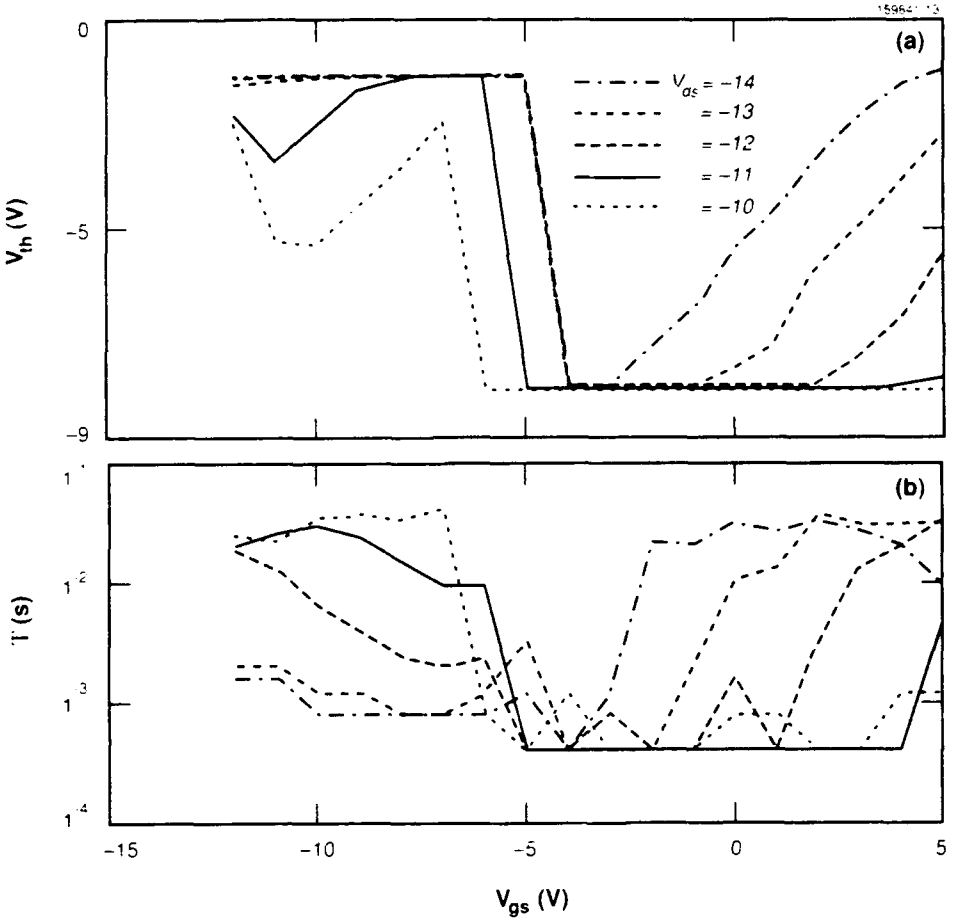



*Figure 13. P-type overlapped gate structure. 1:1 coupling, 4-μm floating gate, 4-μm control gate, 2-μm overlap: (a) maximum threshold voltage and (b) writing time.*

### 2.2.5 Writing N-Type Avalanche Gate Structures

None of the fabricated N-type avalanche-write test structures could be written.

### 2.2.6 Writing P-Type Avalanche Gate Structures

Figure 14 plots an avalanche device with the capacitive coupling ratio of 1:5. The breakdown voltage of the P-drain was around $-18$ V. With a 100 MΩ resistor in series with the programming supply ($V_{pp}$), a current of around 10 nA/V above the breakdown voltage is being injected into the junction. Writing times may not be representative as the capacitance of the probe needle that was used to contact the device terminals and followed the resistor introduced a very long time constant that would not necessarily exist on an integrated circuit. The plots of maximum threshold shift for various programming voltages is linear in gate voltage. This could be very useful in analog charge storage applications. Writing times can be decreased by using more negative programming supply voltages, more positive gate voltages, or both. Maximum threshold shifts can be be achieved with programming supply voltages as low as $-22$ V.

Figures 15 and 16 represent data taken from similar devices as described above, but with capacitive coupling ratios of 1:2 and 1:1, respectively. Again, the maximum thresholds are linear in gate voltage. Programming time increases with increasing coupling ratios. The relative spacing between separate $V_{pp}$ curves for a given $V_{gs}$ also decreases with increasing coupling, as seen earlier in the P-type stacked gate structures.

### 2.2.7 Erasing

Capacitor structures were fabricated along with the floating gate test structures in order to characterize the interpoly and channel dielectrics. Figure 17 shows the I-V characteristics of the capacitor dielectrics. Detectable current begins to flow in the interpoly dielectric due to Fowler-Nordheim tunneling at around 12 V and in the channel dielectric at around 27 V. This difference in tunneling voltage is probably due to the presence of asperities on the upper surface of the floating poly layer creating high electric field points. These asperities, then, provide the mechanism whereby reasonable erase voltages may be realized.

Devices were all erased through this interpoly dielectric by applying high positive voltages to the control gate. The plots shown in Figure 18 demonstrate the effects of increasing capacitive coupling on erase characteristics. The three coupling ratios represented are 1:5, 1:2, and 1:1. Higher coupling ratios could not be used, as the tester used to generate the voltages could not exceed 24 V. Two characteristics are quickly identified. First, the point at which erasure begins increases with increased coupling between the control and the floating gates. Second, the range over which intermediate threshold voltages could be achieved increased with increased coupling (i.e., the slope of the threshold versus $V_{gs}$ curve becomes more gradual). Erase times can be decreased by many orders of magnitude by simply increasing the applied erase voltage to the gate. The plots indicate an order-of-magnitude increase in speed for every few volts' increase in gate voltage. Time

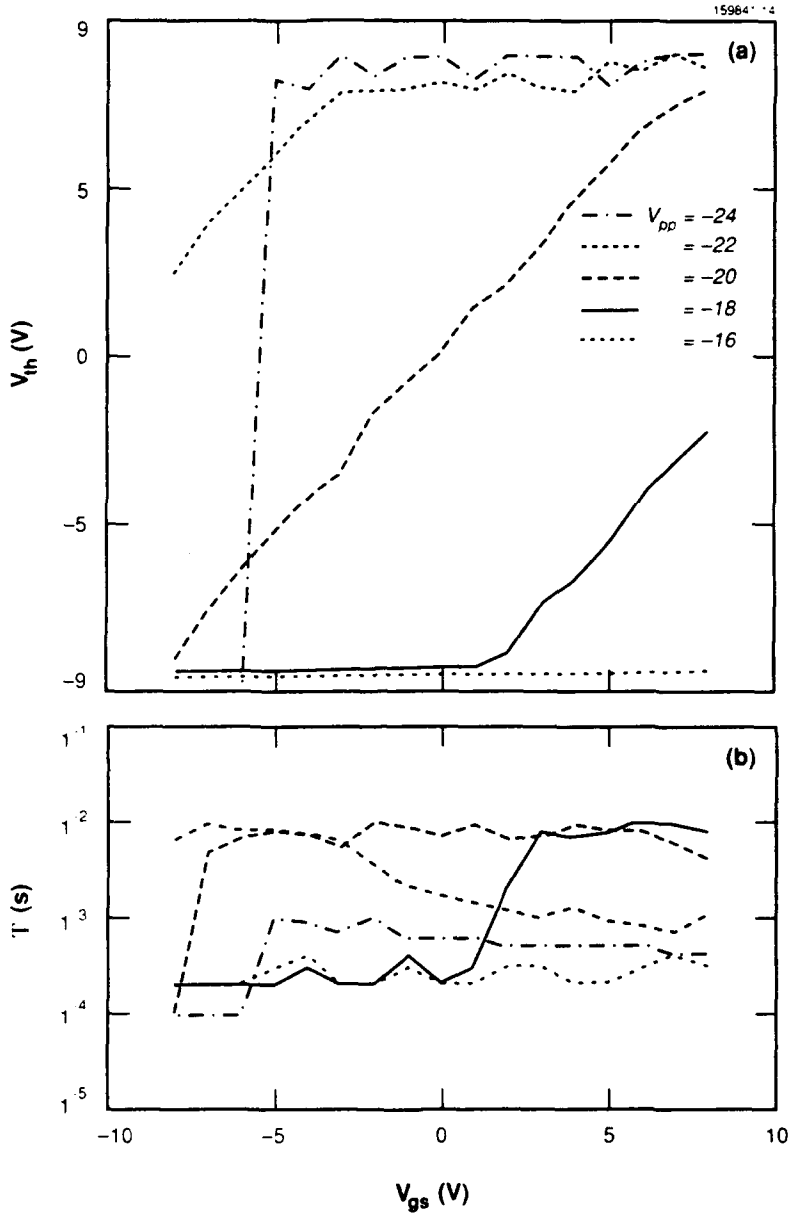



*Figure 14. P-type overlapped gate structure, 1:5 coupling, (a) maximum threshold voltage and (b) writing time.*

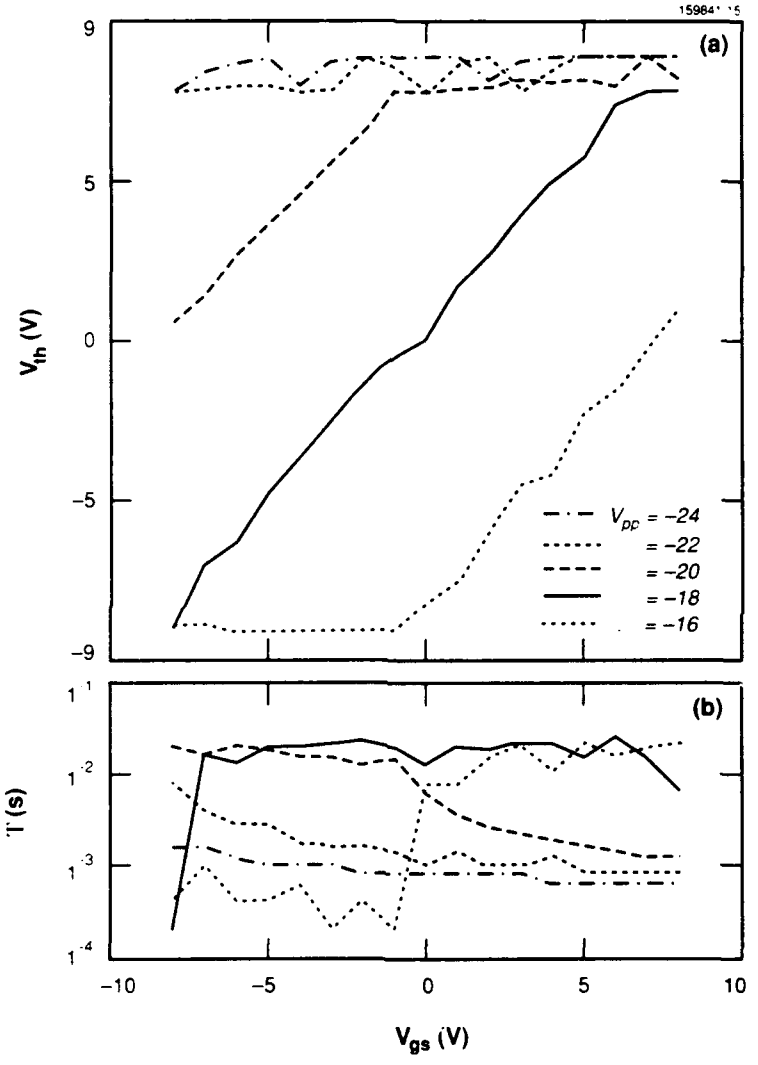



*Figure 15. P-type avalanche structure, 1:5 coupling: (a) maximum threshold voltage and (b) writing time.*

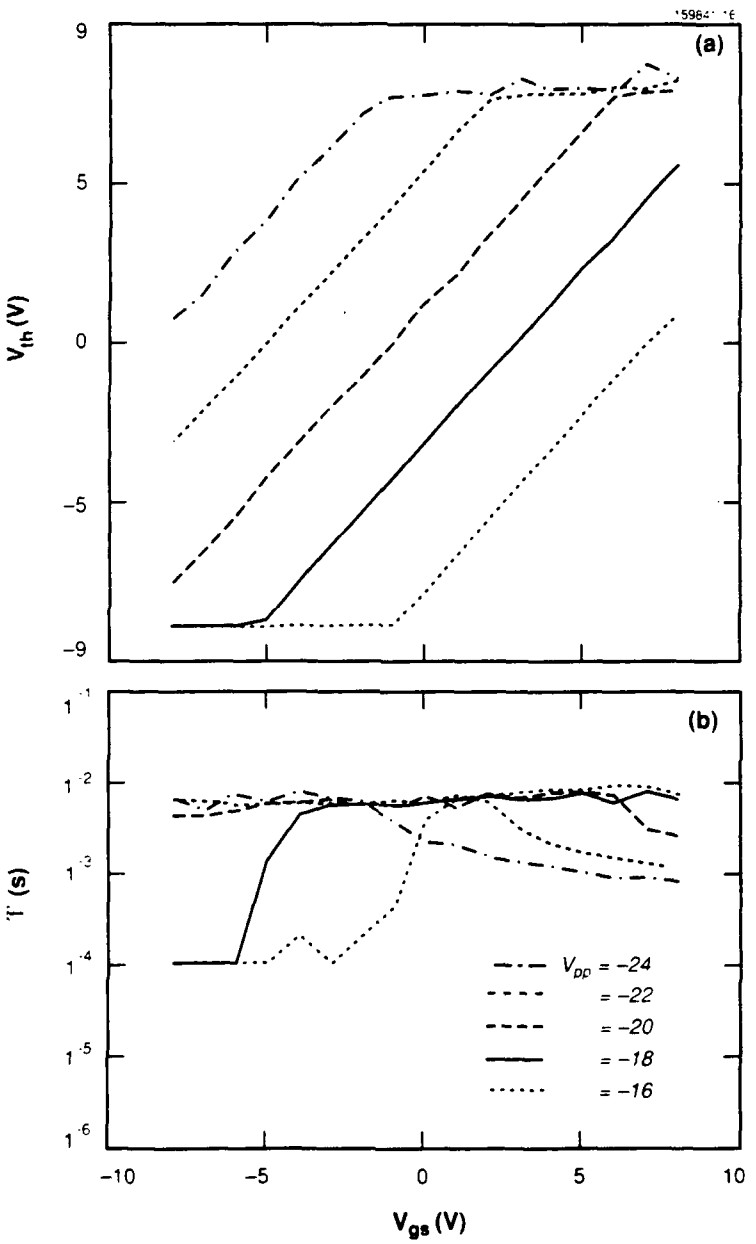



*Figure 16. P-type avalanche structure, 1:1 coupling: (a) maximum threshold voltage and (b) writing time.*

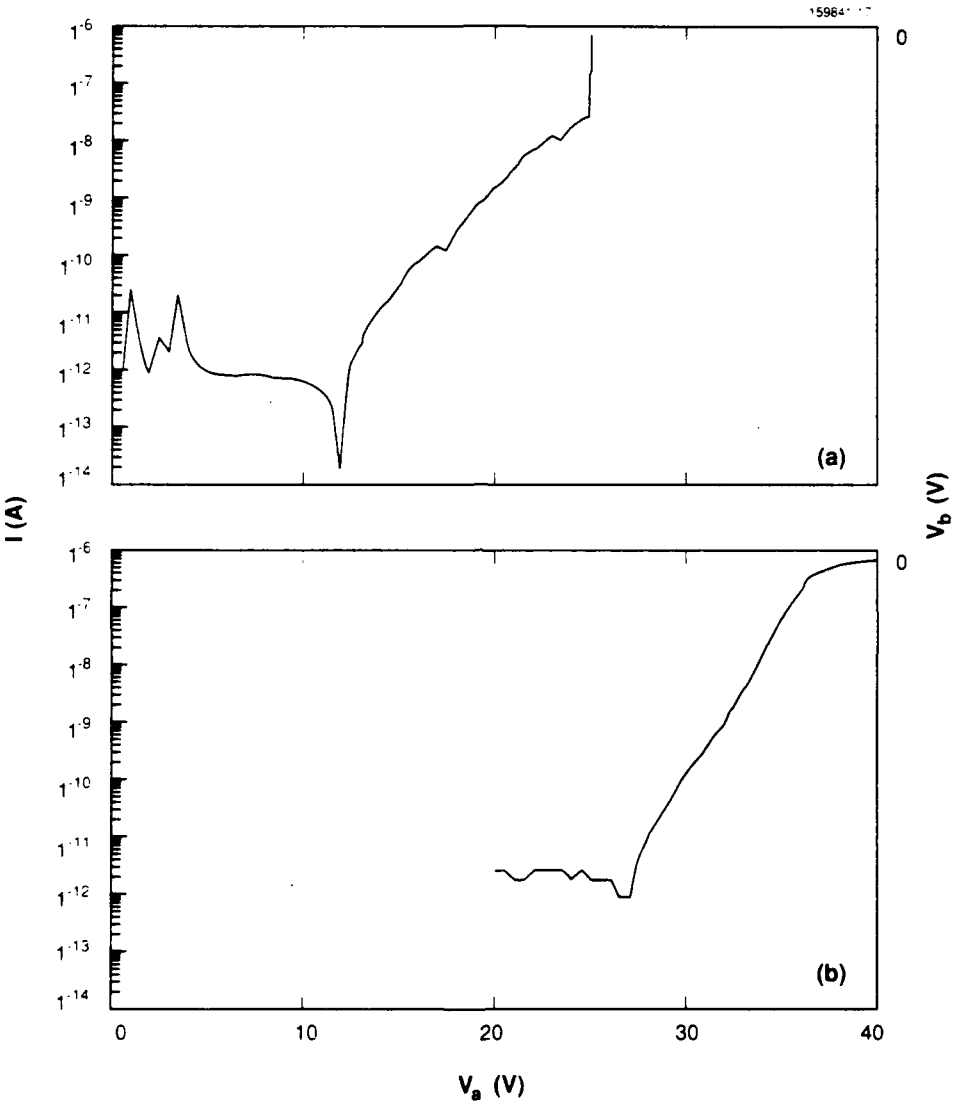



*Figure 17. I-V characteristics of capacitor dielectrics: (a) interpoly and (b) channel.*

plots are limited to the time of a single erase pulse. This represents the minimum absolute time obtainable for erasing and, consequently, also represents the resolution of the time representation.

### 2.2.8 Window Closure

A very important parameter in EEPROM devices is the change in the threshold window as a function of the number of write/erase cycles. The threshold window is defined as the magnitude of the difference between the written and erased threshold voltages achieved under a set of static operating conditions. During both write and erase, electrons travel through one of the oxides, and a fraction of them are trapped. These permanently trapped charges create a field that opposes the transport of electrons to or from the floating gate. Under static operating conditions, this results in a decreased threshold shift as the traps in the oxide are filled.

Figure 19 shows the cycle data taken on a stacked gate P-type structure with a coupling ratio of 1:2. Figure 19(a) shows the resultant write (upper trace) and erase (lower trace) threshold voltages reached on each cycle. Figure 19(b) plots the difference between these two traces, recording the actual threshold window as a function of write/erase cycles. The first few cycles involved setting up the conditions to be used during the cycling operation and can be ignored. The threshold window starts off at 6 V and closes to just over 2 V in 1000 cycles. This is obviously a rapid window closure in comparison with commercial EEPROM technology. This is, in part, due to the substantially thicker oxides through which the electrons are tunneling.

However, constant current stress measurements of capacitor structures fabricated on the same die indicate that the oxides have inherently high trap densities. Figure 20 illustrates the effects of Fowler-Nordheim electron injection on the upper and lower oxides. A constant current is forced through the dielectrics over a period of many minutes and the voltage necessary to maintain this current is recorded. The fluence is the total number of electrons per unit area that have passed through the oxide. The resultant voltage shift necessary to maintain constant current is a measure of the charge trapped in the oxide. Both the upper and lower oxides exhibit similar electron trap densities. Figure 21 shows the voltage shift versus fluence for an oxide grown in the Group 23 facility at Lincoln Laboratory, which is of comparable thickness ($\approx$ 400 $\AA$). (The rapid rise below $1 \times 10^{13}$ represents charging of capacitance.) The small variation of voltage with fluence in our oxides would indicate that the oxides in the MOSIS process have much higher trap densities than our in-house process. This explains why the MOSIS floating gate structures experience such rapid window closures and demonstrates that modifications to the process might improve this performance considerably.

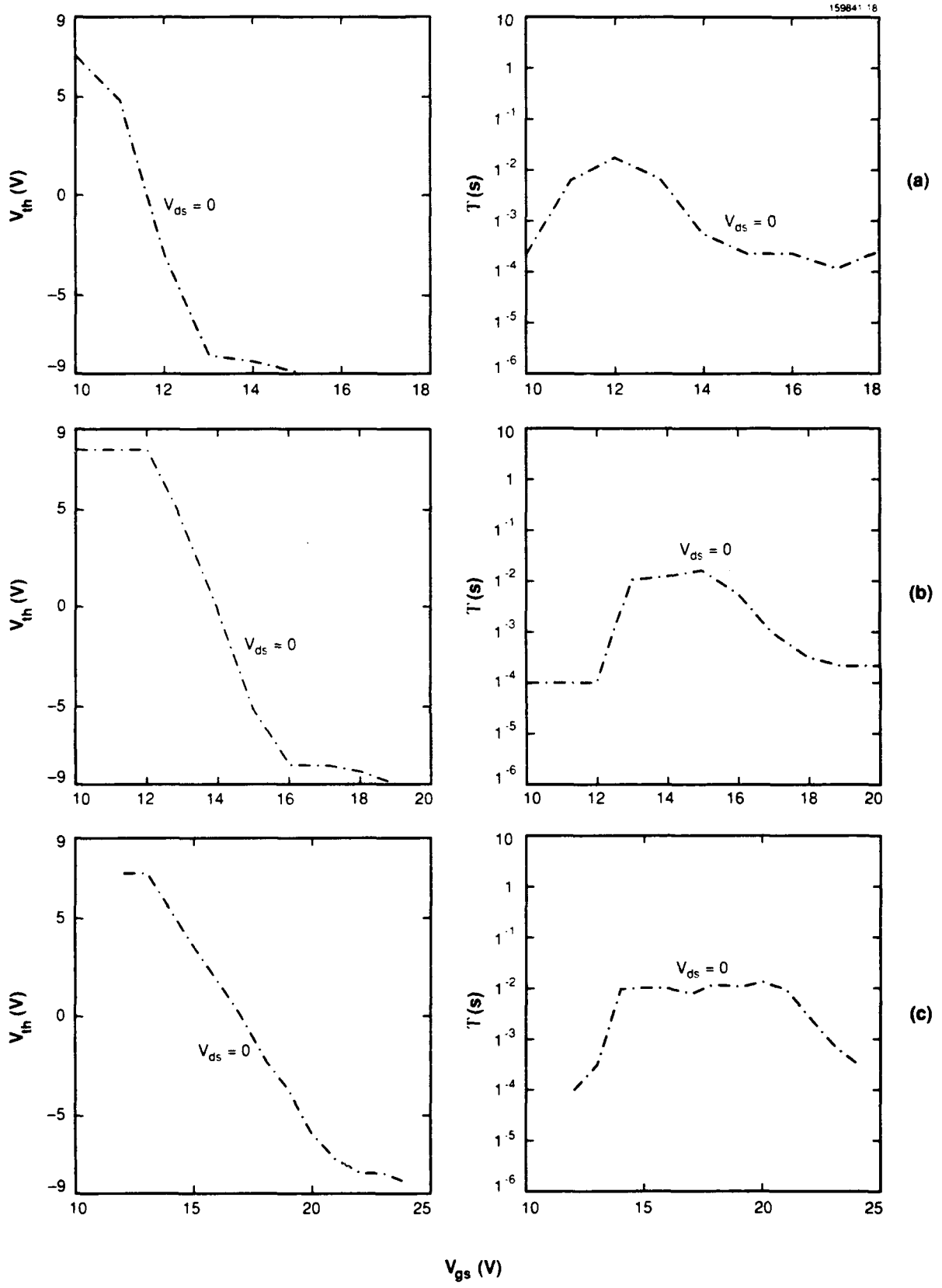



*Figure 18. Erase characteristics for coupling ratios: (a) 1:5, (b) 1:2, and (c) 1:1.*

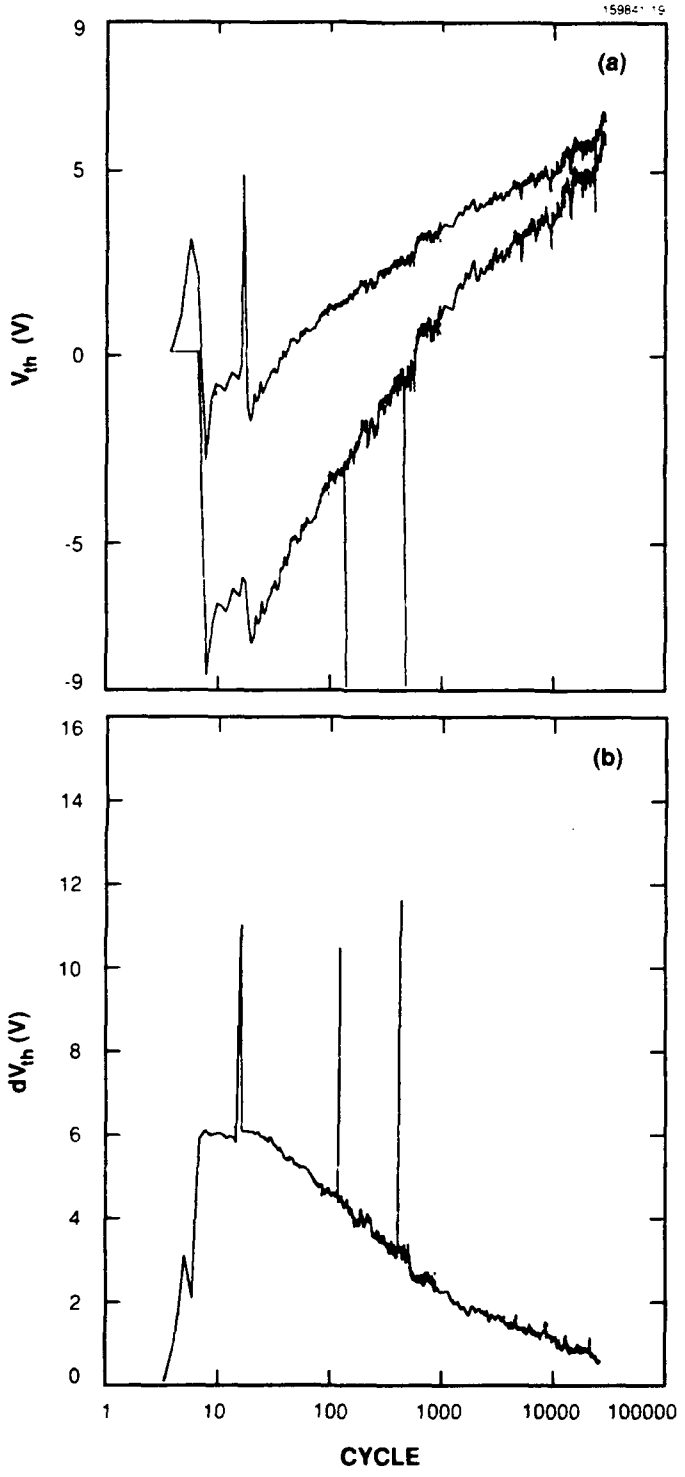



*Figure 19. Floating gate cycle date: (a) write and erase threshold voltages reached and (b) threshold window size.*

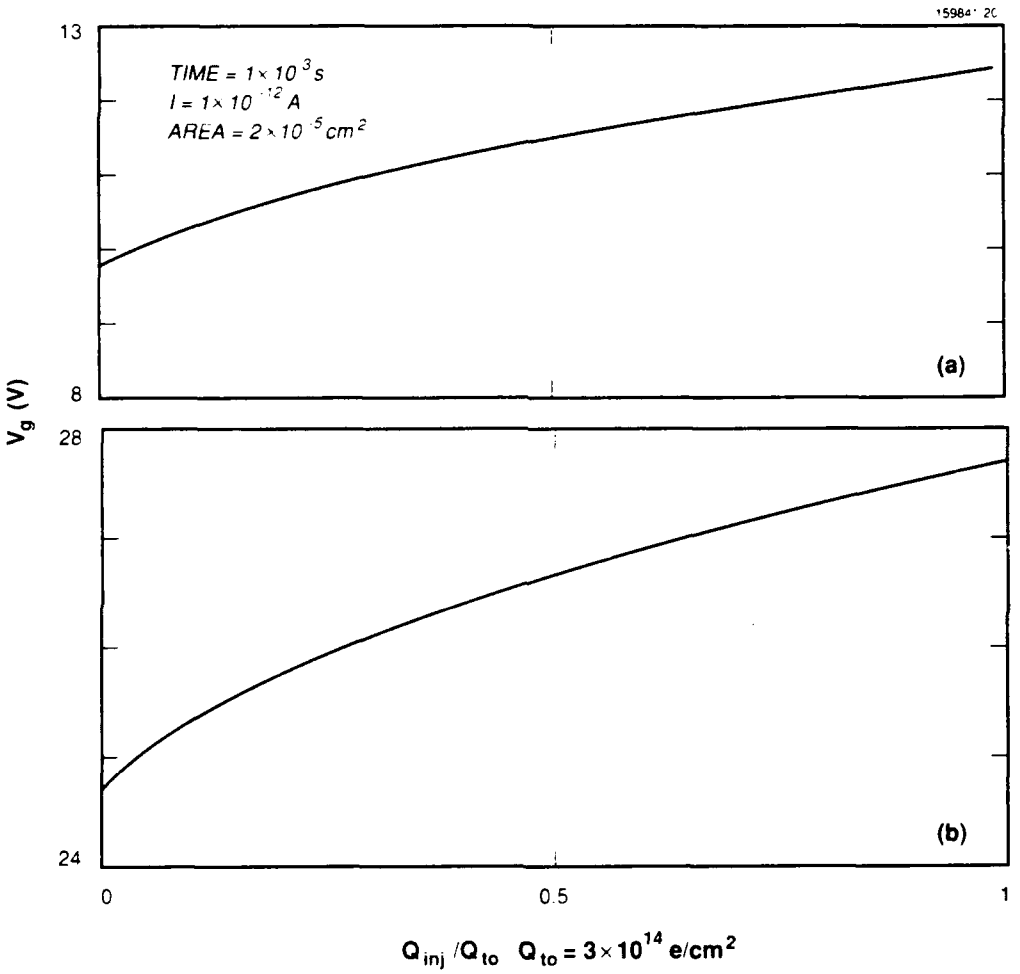



*Figure 20. Constant current stress voltage for low-noise analog dielectrics: (a) interpoly and (b) channel oxides.*

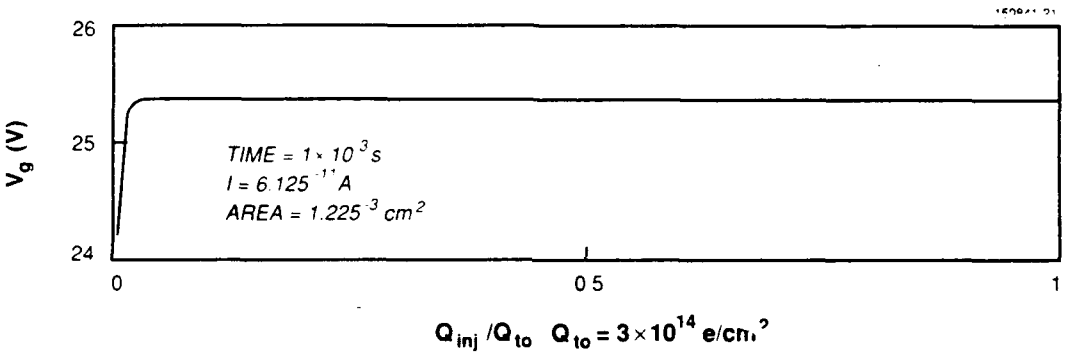



*Figure 21. Constant current stress voltage for in-house dielectric.*

# 3. HIGH-VOLTAGE ADDRESSING CIRCUITRY

In order to selectively route high-voltage programming signals to floating gate structures on a chip, a combination of high-voltage device structures and circuit techniques has been developed. These devices, combined with a source of on-chip high voltage [6] and the floating gate structures just presented, constitute a complete set of circuits for realizing fully integrated EEPROMs.

## 3.1 Method of Testing

Junction breakdown voltages are inversely related to the impurity concentrations of their P-N regions. (However, this is complicated by the addition of fields set up by the gate of an FET around a drain/channel junction.) This implies that using a lower doping concentration for the drains of these devices should extend their voltage handling capacity. This approach was applied to the structures illustrated in Figure 22. The NFET uses the N-well as its drain, and the PFET uses the P-base implant (used in the NPN bipolar transistor). Neither of these layers is masked by the poly during implantation (i.e., they are not self-aligned) and must include an overlap with the thin oxide gate region to ensure channel continuity. In our designs this overlap was 2 $\mu$m with an additional 2 $\mu$m of poly over the field oxide. The contacts to these drains must be surrounded by the lightly doped drain material so as not to come in contact with the surrounding substrate or well. This overlap was 3 $\mu$m in the devices tested and all had 8-$\mu$m channel widths.

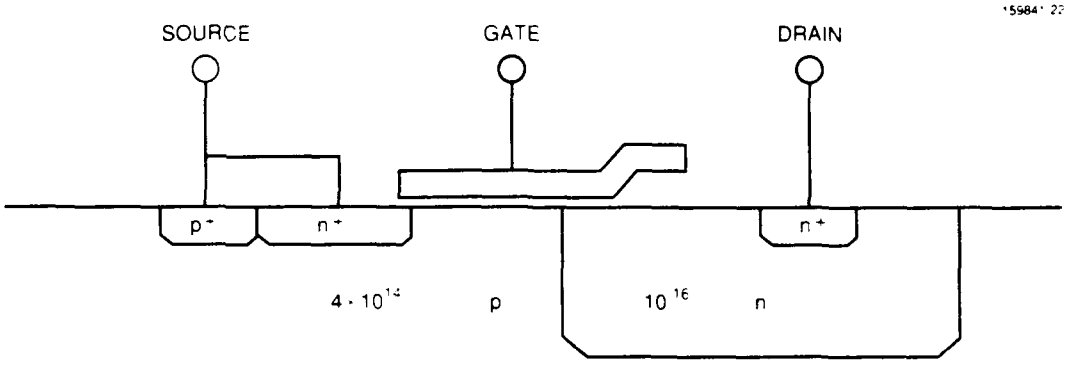
Alone, the N-well has a breakdown voltage around 70 V, and the P-base implant breaks down to the N-well at around 35 V. These voltages are well above those needed; however, the introduction of a gate alters this condition. Testing was performed in two parts. First, the gate was swept over a 30-V range while the drain was stepped over a region of low voltages. Second, the drain was swept over a 30-V range with the gate stepped over a region of low voltages. This technique was selected on the grounds that these devices would normally be powered by an on-chip high-voltage generation circuit that has limited current sourcing capacity and would normally be operating under these conditions. During switching of these devices, connected to such a source, the high voltage is shorted out until it is capable of charging the resultant connected paths back up to high voltages.

In the following discussion, the sources are all connected to ground and the applied voltages are positive or negative with respect to the source depending on device type (N-type, positive; P-type, negative).

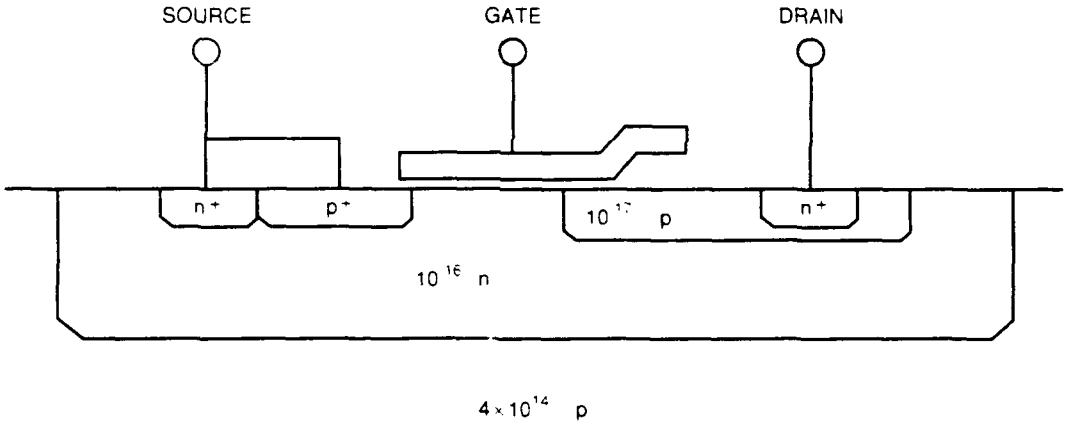
## 3.2 Results

### 3.2.1 High-Voltage NFET

The high-voltage NFET (HVNFET) worked over the full range of voltages as illustrated in Figures 23 and 24. Part (a) of each figure plots the drain current as a function of drain-to-source voltage. Part (b) of each figure is the gate current recorded under the same conditions. The absence




**(a) HIGH VOLTAGE N-CHANNEL FET**




**(b) HIGH VOLTAGE P-CHANNEL FET**

*Figure 22. High-voltage breakdown drain FETs: (a) NFET and (b) PFET.*

of any gate current under these conditions indicates that no breakdown is occurring in the channel region at the drain end and the device is operating as desired.

### 3.2.2 High-Voltage PFET

The high-voltage PFET (HVPFET) suffered a breakdown, probably at the channel surface, at drain voltages of −15 V with zero gate-to-source voltage. This is seen in the corresponding plots given in Figures 25 and 26. Figure 26(b) clearly shows an appreciable gate current below −20-V $V_{ds}$ for all gate voltages tested. Originally, on the fresh device, gate currents were detected at drain voltages as low as −15 V. Charge trapped in the oxide causes the breakdown voltage to experience a "walk out" with continued stress. These plots were made on a device that had already received considerable stress. It is most likely an avalanche breakdown because the gate current indicates that hot carriers are being generated, and the P-base to N-well breakdown voltage is much higher than −15 V without a gate present.

### 3.2.3 HVPFET Fix

This breakdown problem can be overcome by the addition of a separate implant step of lower doping for the HVPFET drain region, or by dropping the voltage across several devices in series. The latter can be accomplished as illustrated in Figure 27. The bias voltage ($V_b$) represents some voltage midway between the upper and lower high-voltage rails. The source of the bias transistor cannot go below the gate voltage of this device so that the upper transistor never sees a drain-to-source potential any greater than the difference between $V_{hi}$ and $V_b$. The rest of the voltage drops across the biasing transistor. This technique could also be extended to any number of bias transistors as needed.

The HVPFET circuit may require more than two transistors to handle voltages above 25 V. At 30 V, gate currents can be detected in both devices, the magnitude of which is determined by the bias voltage. This is illustrated in Figure 28, where the signal gate current ($V_{in}$) and bias gate current ($V_b$) are plotted as a function of bias voltage. The slopes of the two currents have been extended to their point of intersection to identify the optimal operating bias voltage for this arrangement. No detectable currents are generated at this bias potential at programming voltages less than 25 V.

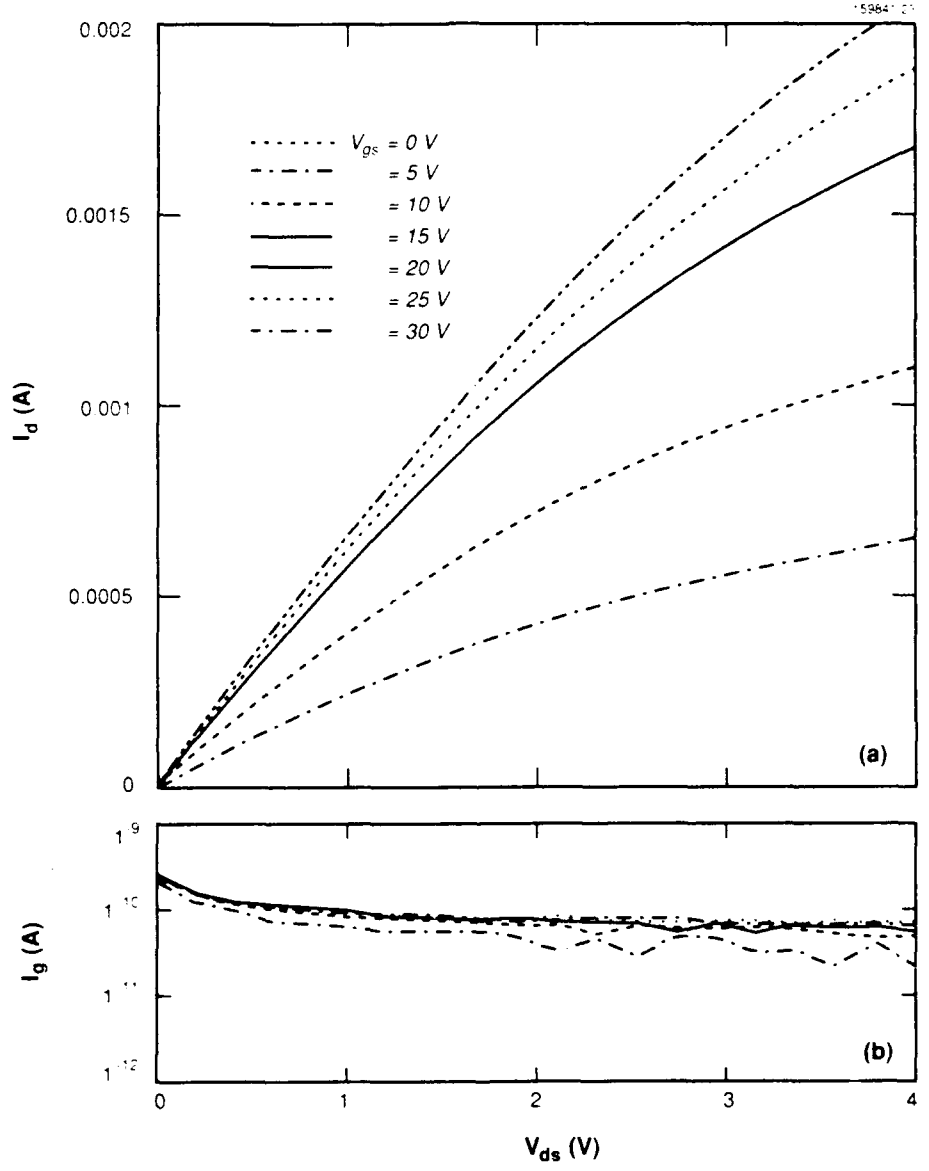



*Figure 23. High-voltage NFET high gate voltage test: (a) drain current versus $V_{ds}$ versus $V_{gs}$ and (b) gate current versus $V_{ds}$ versus $V_{gs}$.*

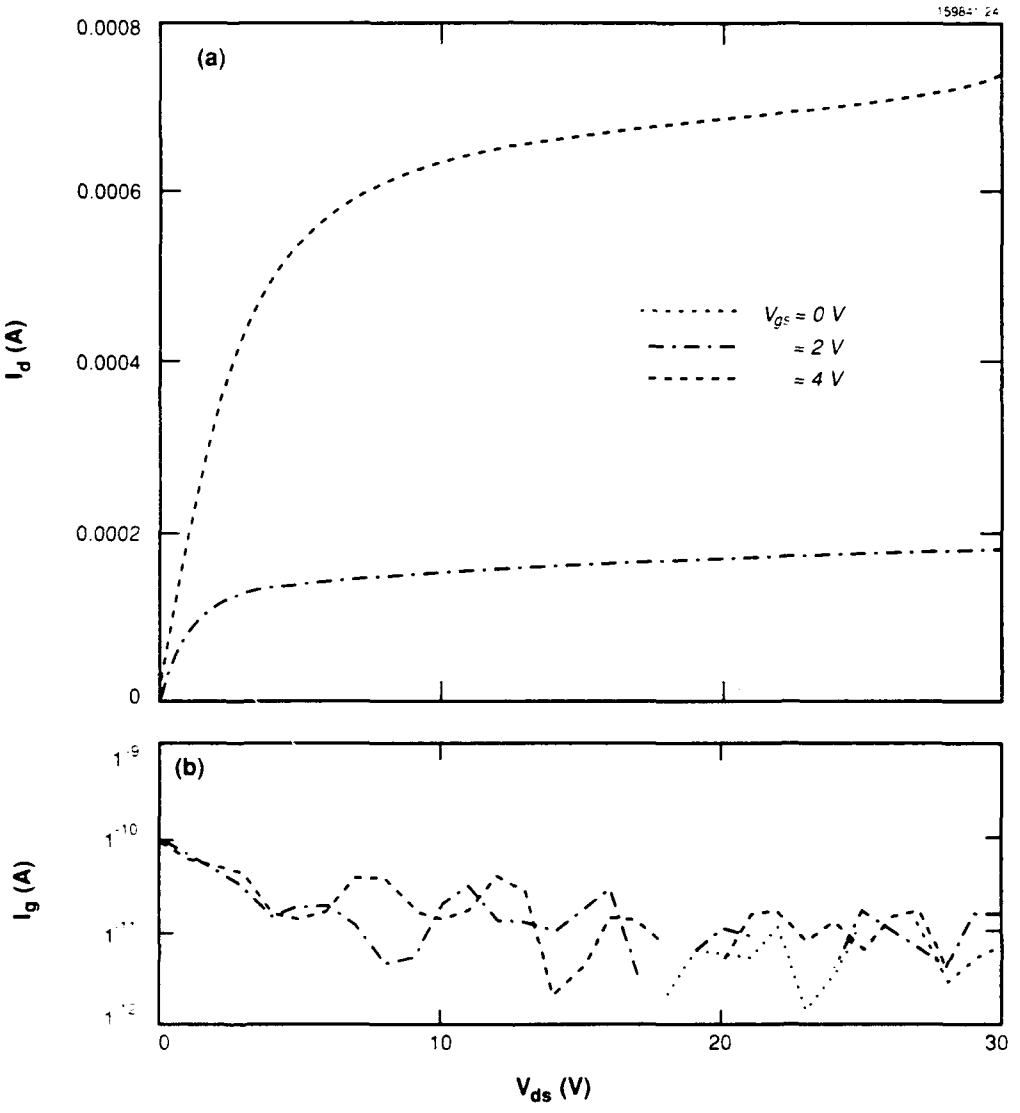



*Figure 24. High-voltage NFET high drain voltage test: (a) drain current versus $V_{ds}$ versus $V_{gs}$ and (b) gate current versus $V_{ds}$ versus $V_{gs}$.*

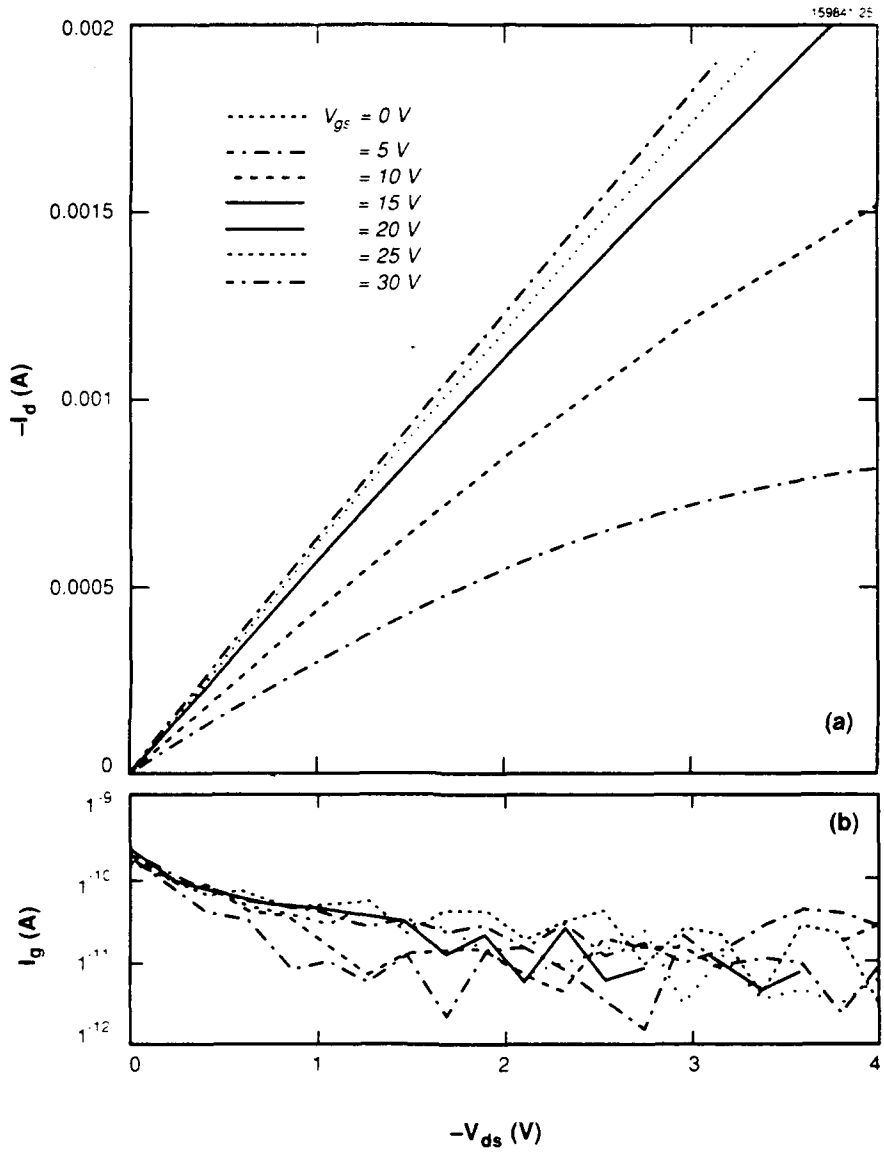



*Figure 25. High-voltage PFET high gate voltage test: (a) drain current versus $V_{ds}$ versus $V_{gs}$ and (b) gate current versus $V_{ds}$ versus $V_{gs}$.*

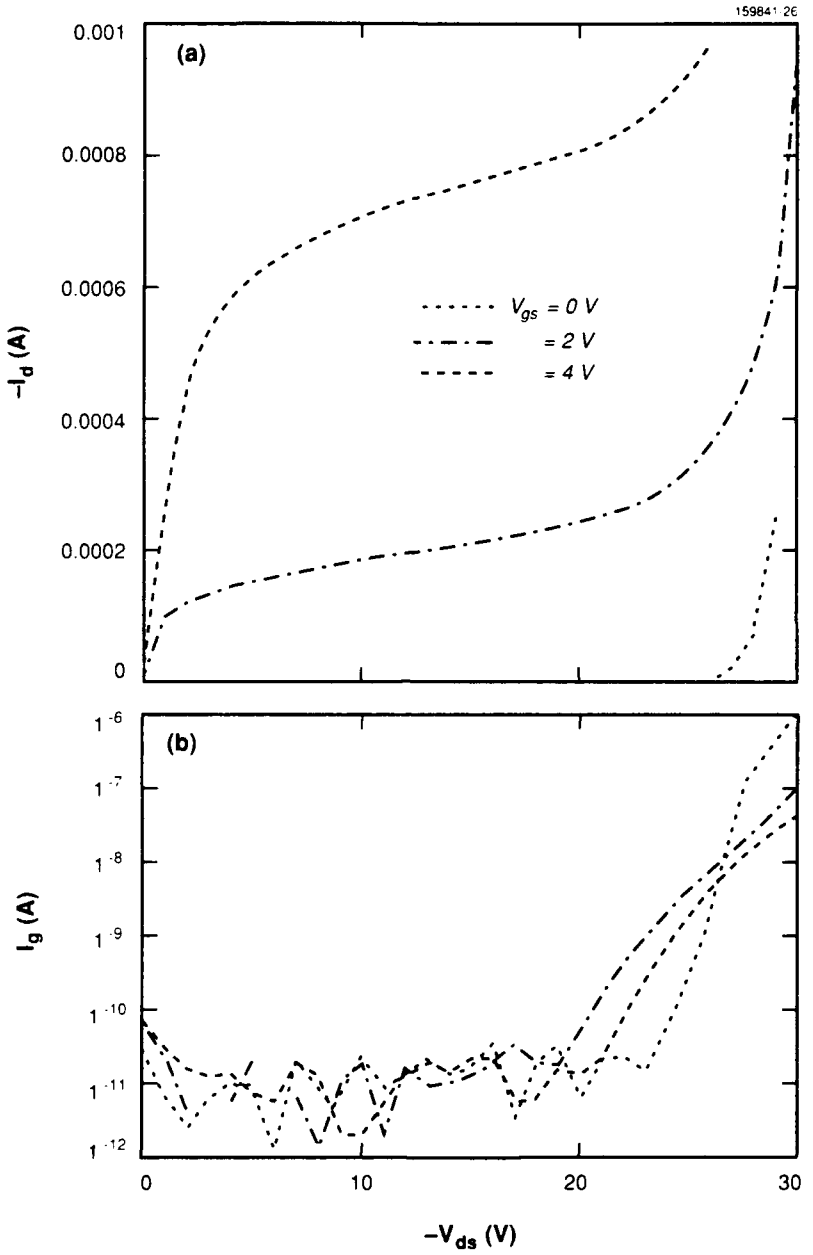



*Figure 26. High-voltage PFET high drain voltage test. The breakdown voltage on this device has "walked out" from the original −15 V to about −24 V. (a) Drain current versus $V_{ds}$ versus $V_{gs}$ and (b) gate current versus $V_{ds}$ versus $V_{gs}$.*

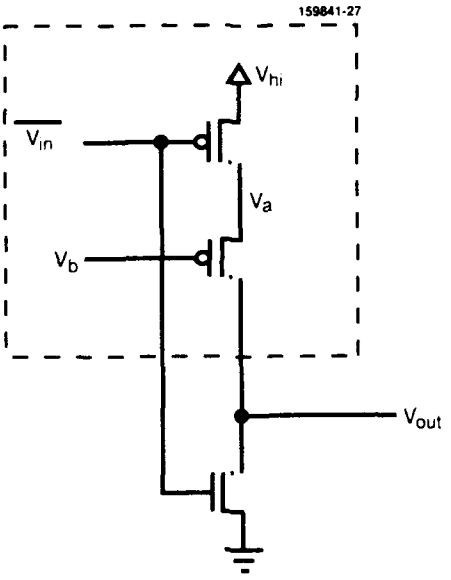



*Figure 27. High-voltage PFET divider circuit.*

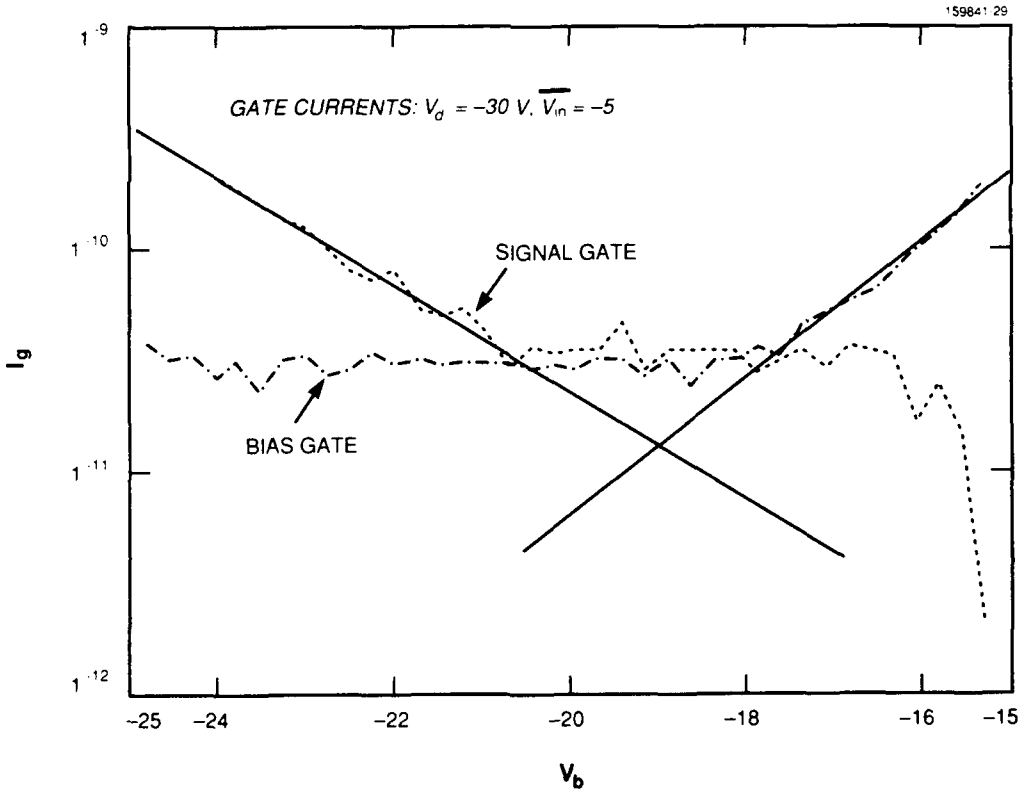



*Figure 28. Signal and bias gate currents as a function of bias voltage. The intercept of the slope extensions identifies the optimal bias potential.*

# 4. EEPROM CIRCUITS

## 4.1 A High-Voltage Addressing Circuit

These high-voltage transistors can be included in a high-voltage addressing circuit as shown in Figure 29. The select signals are conventional TTL-level signals generated by an address decoder typically used in memory cells. Cells including the HVPFET circuit and the high-voltage addressing circuit are also being fabricated.

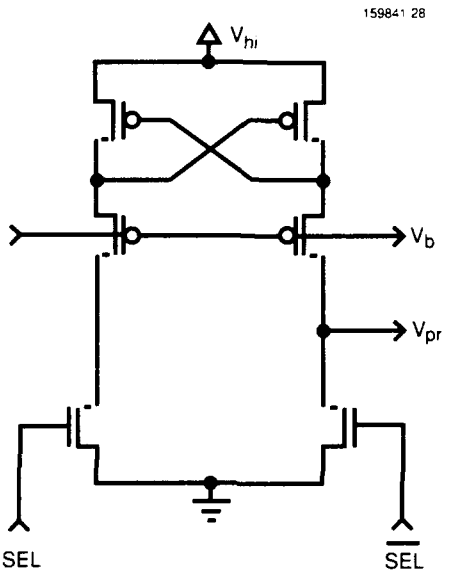



*Figure 29. High-voltage addressing circuit.*

## 4.2 EEPROM Core Circuit

Also in fabrication are the combined high-voltage elements of a complete EEPROM. This includes an $8 \times 8$ array of P-type stacked floating gate structures with series select transistors and high voltage addressing circuits for the array. No decoding or on-chip high-voltage generation circuitry has been included. The idea for this first pass is to begin to identify the optimal combination of parts for an integrated EEPROM. For the moment, the charge pump will be emulated by including a high resistance in series with the high-voltage supply.

# 5. CONCLUSION

All the parts needed to build a self-contained EEPROM capability using the low-noise analog CMOS process available through MOSIS have been fabricated and characterized. A first pass EEPROM core is in fabrication for the purpose of investigating issues involving integration of the different subcircuit parts.

One of the main problems facing this technology is the rapid threshold window closure. This limits the applicability of this technology to applications that require infrequent reprogramming. This would include analog trimming, EPLDs, and some neural network applications. This problem can be overcome with adaptive programming techniques that employ some type of feedback, such as interrogating and rewriting the memory contents until they return correct. Competitive neural networks may also not suffer as greatly from window closure as long as the stored charge increases or decreases with write or erase cycles as they are thought to be somewhat insensitive to absolute weight adjustment magnitude.

Conversations with the fabricator also indicate that rather minor changes to the current process may reduce the trap densities of the oxides, though there is no way to ensure that such changes will be made. This brings up a final point: all these devices were characterized using the existing MOSIS process. Any changes by the foundry, such as reducing the first poly surface asperities to reduce interpoly tunneling for switched capacitor applications, can and will alter these findings at any time and without notification.

When a fully integrated EEPROM is finally worked out, these cells will be made available as a library of parts from which various sized memories can be built. A preliminary Magic technology file can also be made available though, at present, it does not include design rule checking for any of the new layers and it requires some changes to the Magic "include" files. CIF files for any of the test devices are available from mann@vlsi.ll.mit.edu.

## REFERENCES


1. E. Sackinger and W. Guggenbuhl, "An analog trimming circuit based on a floating-gate device," *IEEE J. of Solid-State Circuits*, **23**, 1437–1440 (1988).

2. M. Holler, S. Tam, H. Castro, and R. Benson, "An electrically trainable artificial neural network (ETANN) with 10240 'floating gate' synapses," *Proc. Int. Joint Conf. on Neural Networks*, II-191–196 (1989).

3. A. Kramer, K.C. Sin, B. Gupta, R. Chu, and P.K. Ko, "A highly compact linear weight function based on the use of EEPROMs," Neural Networks for Computing Conference, Snowbird, Utah (1990).

4. L. Glasser, "A UV write-enabled PROM," Chapel Hill Conference on VLSI, 61–65 (1985).

5. G. Samachisa, C. Su, Y. Kao, G. Smarandoiu, C.M. Wang, T. Wong, and C. Hu, "A 128-K flash EEPROM using double-polysilicon technology," *IEEE J. Solid-State Circuits*, **22**, 676–682 (1987).

6. J.F. Dickson, "On-chip high-voltage generation in MNOS integrated circuits using an improved voltage multiplier technique," *IEEE J. Solid-State Circuits*, **11**, 374–378, (1976).

# REPORT DOCUMENTATION PAGE

*Form Approved*
*OMB No. 0704-0188*



| 1. AGENCY USE ONLY *(Leave blank)* | 2. REPORT DATE | 3. REPORT TYPE AND DATES COVERED |
|---|---|---|
| | 1 November 1990 | Technical Report |

**4. TITLE AND SUBTITLE**

Floating Gate Circuits in MOSIS

**5. FUNDING NUMBERS**

C — F19628-90-C-0002
PE — 62301E
PR — 335

**6. AUTHOR(S)**

James R. Mann

**7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)**

Lincoln Laboratory, MIT
P.O. Box 73
Lexington, MA 02173-9108

**8. PERFORMING ORGANIZATION REPORT NUMBER**

TR-824

**9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)**

Defense Advanced Research Projects Agency
1400 Wilson Boulevard
Arlington, VA 22209

**10. SPONSORING/MONITORING AGENCY REPORT NUMBER**

ESD-TR-90-114

**11. SUPPLEMENTARY NOTES**

None

**12a. DISTRIBUTION/AVAILABILITY STATEMENT**

Approved for public release; distribution is unlimited.

**12b. DISTRIBUTION CODE**

**13. ABSTRACT** *(Maximum 200 words)*

The MOSIS foundry offers a double poly CMOS process that can be used as a floating gate technology, albeit, not with the same performance as commercial EEPROM foundries. This report characterizes the structures and programming techniques necessary to build floating gate structures and associated high voltage addressing circuitry on the low noise analog process available through MOSIS. Techniques that are used include Fowler-Nordheim tunneling, channel hot-electron injection, and avalanche injection. The dielectric materials between the floating gate and both the control gate and substrate are characterized. Unconventional lightly doped drain FET devices and additional circuit techniques for handling the high voltage programming signals are presented.

**14. SUBJECT TERMS**

electrically erasable programmable read-only memory (EEPROM)
floating gate
Fowler-Nordheim tunneling
channel hot-electron injection
avalanche injection

**15. NUMBER OF PAGES**
54

**16. PRICE CODE**

| 17. SECURITY CLASSIFICATION OF REPORT | 18. SECURITY CLASSIFICATION OF THIS PAGE | 19. SECURITY CLASSIFICATION OF ABSTRACT | 20. LIMITATION OF ABSTRACT |
|---|---|---|---|
| Unclassified | Unclassified | Unclassified | SAR |